NISTIR 8271

Face Recognition Vendor Test (FRVT)
Part 2: Identification

Patrick Grother
Mei Ngan
Kayee Hanaoka

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8271

National Institute of
Standards and Technology
U.S. Department of Commerce

NISTIR 8271

Face Recognition Vendor Test (FRVT)
Part 2: Identification

Patrick Grother
Mei Ngan
Kayee Hanaoka
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8271

September 2019

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Walter Copan, NIST Director and Undersecretary of Commerce for Standards and Technology

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

National Institute of Standards and Technology Interagency or Internal Report 8271
Natl. Inst. Stand. Technol. Interag. Intern. Rep. 8271, 186 pages (September 2019)

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8271

ACKNOWLEDGMENTS

The authors are grateful to Wayne Salamon and Greg Fiumara at NIST for designing robust software infrastructure for image and template storage and parallel execution of algorithms across our computers. Thanks also to Brian Cochran at NIST for providing highly available computers and network-attached storage.

DISCLAIMER

Specific hardware and software products identified in this report were used in order to perform the evaluations described in this document. In no case does identification of any commercial product, trade name, or vendor, imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

Executive Summary

This report updates and extends NIST Interagency Report 8238, documenting the evaluation of automated face recognition algorithms submitted to NIST in November 2018. The algorithms, which implement one-to-many identification of faces appearing in two-dimensional images, are prototypes from the research and development laboratories of mostly commercial suppliers, and are submitted to NIST as compiled black-box libraries implementing a NIST-specified C++ test interface. The report therefore does not describe how algorithms operate.

The evaluation used four datasets - frontal mugshots, profile views, webcam photos and wild images - and the report lists accuracy results alongside developer names. It will therefore be useful for comparison of face recognition algorithms and assessment of absolute capability. The primary dataset is comprised of 26.6 million reasonably well-controlled live portrait photos of 12.3 million individuals. The three smaller datasets contain more unconstrained photos: 3.2 million webcam images; 200 thousand side-view images; and 2.5 million photojournalism and amateur photographer photos. These datasets are sequestered at NIST, meaning that developers do not have access to them for training or testing. The last dataset, however, consists of images drawn from the internet for testing purposes so while it is not truly sequestered, its composition is unknown to the developers.

The evaluation was run in three phases, starting Feburary, June, and November 2018 respectively, with developers receiving technical feedback between phases. Results for 127 algorithms from 41 developers were published in November 2018 as NIST Interagency Report 8238. This update adds results for an additional 76 algorithms from 42 developers submitted in October 2018. At that time seven developers ceased participation, and nine developers started. The developer totals constitute a substantial majority of the face recognition industry.

The major result given in NIST IR 8238 was that massive gains in accuracy have been achieved in the last five years (2013-2018) and these far exceed improvements made in the prior period (2010-2013). While the industry gains were broad - at least 30 developers' algorithms outperformed the most accurate algorithm from late 2013 - there remains a wide range of capability. While this report shows accuracy gains only over the course of 2018, the most accurate algorithm reported here is substantially more accurate than anything reported in NIST IR 8238. This is evidence that face recognition development continues apace, and that FRVT reports are but a snapshot of contemporary capability.

From discussion with developers, the accuracy gains stem from the adoption of deep convolutional neural networks. As such, face recognition has undergone an industrial revolution, with algorithms increasingly tolerant of poorly illuminated and other low quality images, and poorly posed subjects. One related result is that a few algorithms correctly match side-view photographs to galleries of frontal photos, with search accuracy approaching that of the best c. 2010 algorithms executing frontal-frontal search. The capability to recognize under a 90-degree change in viewpoint - pose invariance - has been a long-sought milestone in face recognition research.

With good quality portrait photos, the most accurate algorithms will find matching entries, when present, in galleries containing 12 million individuals, with rank one miss rates of approaching 0.1%. The remaining errors are in large part attributable to long-run ageing, facial injury and poor image quality. In at least 5% of images identification often succeeds (i.e. the mate is returned at rank 1) but recognition similarity scores are weak such that true and false matches become indistinguishable, and human adjudication becomes necessary.

From Fall 2019 this report will be updated continuously as new algorithms are submitted to FRVT, and run on new datasets. Participation in the one-to-many identification track requires a devloper to first demonstrate high accuracy in the one-to-one verification track of FRVT.

Scope and Context

Audience: This report is intended for developers, integrators, end users, policy makers and others who have some familiarity with biometrics applications. The methods and metrics documented here will be of interest to organizations engaged in tests of face recognition algorithms. Some of these have been incorporated in the ISO/IEC 19795 Part 1 Biometric Testing and Reporting Framework standard, now under revision.

Prior benchmarks: Automated face recognition accuracy has improved massively in the two decades since initial commercialization of the various technologies. NIST has tracked that improvement through its conduct of regular independent, free, open, and public evaluations. These have fostered improvements in the state of the art. This report serves as an update to the NIST Interagency Report 8238 on performance of face identification algorithms, published in November 2018.

Scope: As with NIST IR 8238, this report documents recognition results for four databases containing in excess of 30.2 million still photographs of 14.4 million individuals. This constitutes the largest public and independent evaluation of face recognition ever conducted. It includes results for accuracy, speed, investigative vs. identification applications, scalability to large populations, use of multiple images per person, images of cooperative and non-cooperative subjects.

The report also includes results for ageing, recognition of twins, and recognition of profile-view images against frontal galleries. It otherwise does not address causes of recognition failure, neither image-specific problems nor subject-specific factors including demographics. Separate reports on demographic dependencies in face recognition will be published in the future. Additionally out of scope are: performance of live human-in-the-loop transactional systems like automated border control gates; human recognition accuracy as used in forensic applications; and recognition of persons in video sequences (which NIST evaluated separately [9]). Some of those applications share core matching technologies that *are* tested in this report.

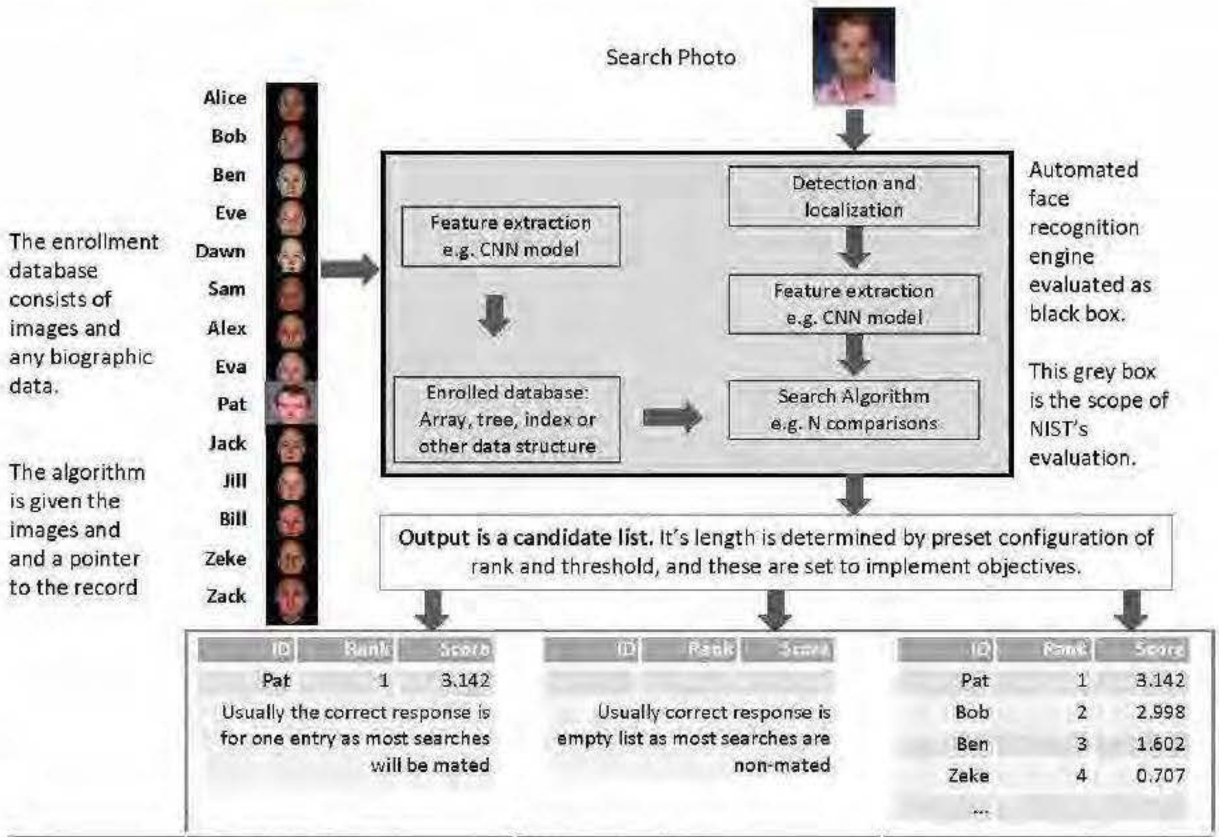
Images: Three kinds of images are employed. The primary dataset is a set of law enforcement mugshot images (Fig. 3) which are enrolled and then searched with three kinds of images: 1) other mugshots (i.e. within-domain); 2) profile-view photographs (90 degree cross-view); 3) lower quality webcam images (Fig. 4) collected in similar detention operations (cross-domain); Additionally wild images (Fig. 6) are searched against other wild images.

Participation and industry coverage: The report includes performance figures for 203 prototype algorithms from the research laboratories of 51 commercial developers and one university. This represents a substantial majority of the face recognition industry, but only a tiny minority of the academic community. Participation was open worldwide. While there is no charge for participation, developers incur some software engineering expense in implementing their algorithms behind the NIST application programming interface (API). The test is a black-box test where the function of the algorithm, and the intellectual property associated with it, is hidden inside pre-compiled libraries.

Recent technology development: Most face recognition research with deep convolutional neural networks (CNNs) has been aimed at achieving invariance to pose, illumination and expression variations that characterize photojournalism and social media images. The initial research [18, 24] employed large numbers of images of relatively few ($\sim 10^4$) individuals to learn invariance. Inevitably much larger populations ($\sim 10^7$) were employed for training [11, 20] but the benchmark, Labeled Faces in the Wild with (essentially) an equal error rate metric [12], represents an easy task, one-to-one verification at very high false match rates. While a larger scale identification benchmark duly followed, Megaface [15], its primary metric, rank one hit rate, contrasts with the high threshold discrimination task required in most large-population applications of face recognition, namely credential de-duplication, and background checks. There, identification in galleries containing up to 10^8 individuals must be performed using a) very few images per individual and b) stringent thresholds to afford very low false positive identification rates. FRVT 2018 was launched to measure the capability of the new technologies, including in these two cases. FRVT has included open-set identification tests since 2002, reporting both false negative and positive identification rates [7].

Performance metrics for applications: This report documents the performance of one-to-many face recognition algorithms. The word "performance" here refers to recognition accuracy and computational resource usage, as measured by executing those algorithms on massive sequestered datasets.

This report includes extensive tabulation of recognition error rates germane to the main use-cases for face search technology. The Figure below, inspired by the Figure 1 in [25] differentiates different applications of the technolgy. The last row directs readers to the main tables relevant to those applications, respectively threshold-based and rank-based metrics that are special cases of the metrics given in section 3. The terms negative identification and positive identification are taken from the ISO/IEC 2382-37:2017 standardized biometrics vocabulary.



	Positive identification	Negative identification	Post-event investigation
Example	Access to a gym or cruise ship	Watchlist e.g. Detection of deportee or duplicate drivers license applications	Crime scene photos, or of detainee without ID documents.
Claim of identity	Implicit claim to be enrolled	Implicit claim to not be enrolled	No claim: Inquiry
Threshold	High, to implement security objective	High, to limit false positives	Zero
Num. candidates	1	0	L, set by request to algorithm
Human role	Review candidate to assist user in resolution of false negatives, or to detect impostor	Review candidate to determine false positive or correct hit	Review multiple candidates, refer possible hits to examiners see [26]
Intended human involvement frequency	Rare – approx. the false negative identification rate plus prior probability of impostor	Rare – approx. the false positive identification rate plus prior probability of an actual mate	Always
Performance metric of interest	**FNIR at low FPIR. See sec. 3.1, 3.2 and Tables 10, 19**	**FNIR at low FPIR. See sec. 3.1, 3.2 and Tables 10, 19**	**FNIR at ranks 1... 50, say. FPIR = 1. See sec. 3.2 and Table 12, 14, 16**

The algorithms are specifically configured for these applications by setting thresholds and candidate list lengths. Both

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

rank-based metrics and threshold-based metrics include tradeoffs. In investigation, overall accuracy will be reduced if labor is only available to review a few candidates from the automated system. Note that when a fixed number of candidates are returned, the false positive identification rate of the automated face recognition engine will be 100%, because a probe image of anyone not enrolled will still return candidates. In identification applications where false positives must be limited to satisfy reviewer labor availability or a security objective, higher false negative rates are implied. This report includes extensive quantification of this threshold-based tradeoff. See Sec. 3

Template diversity: The FRVT is designed to evaluate black-box technologies with the consequence that the templates that hold features extracted from face images are entirely proprietary opaque binary data that embed considerable intellectual property of the developer. Despite migration to CNN-based technologies there is no consensus on the optimal feature vector dimension. This is evidenced by template sizes ranging from below 100 bytes to more than four kilobytes. This diversity of approaches, suggests there is no prospect of a standard template something that would require a common feature set to be extracted from faces. Interoperability in automated face recognition remains solidly based on images and documentary standards for those, in particular the ICAO portrait [29] specification deriving from the ISO/IEC 19794-5 Token frontal [26] standard, which are similar to certain ANSI/NIST Type 10 [28] formats.

Training: The algorithms submitted to NIST have been developed using image datasets that developers do not disclose. The development will often include application of machine learning techniques and will additionally involve iterative training and testing cycles. NIST itself does not perform any training and does not refine or alter the algorithm in any way. Thus the model, data files, and libraries that define an algorithm are fixed for the duration of the tests. This reflects typical operational reality where recognition software, once installed, is fixed and constant until upgraded. This situation persists because on-site training of algorithms on customer data is atypical essentially because training is not a turnkey process.

Automated search and human review: Virtually all applications using automated face search require human review of the outputs at some frequency: Always for investigational applications; rarely in positive identification applications, after rejection (false or otherwise); and rarely in negative identification applications, after an alarm (false or otherwise). The human role is usually to compare a reference image with the query image or the live-subject if present, to render either a definitive decision on "exclusion" (different subjects), or "identification" (same subject), or a declaration that one or both images have "no value" and that no decision can be made. Note that automated face recognition algorithms are not built to do exclusion - low scores from a face comparison arise from different faces *and* poor quality images of the same face.

Human reviewers make recognition errors [5, 19, 27] and are sensitive to image acquisition and quality. Accurate human review is supported by high resolution - as specified in the Type 50, 51 acquisition profiles of the ANSI/NIST Type 10 record [28], and by multiple non-frontal views as specified in the same standard. These often afford views of the ear. Organizations involved in image collection should consider supporting human adjudication by collecting high-resolution frontal and non-frontal views, preparing low resolution versions for automated face recognition [26], and retaining both for any subsequent resolution of candidate matches. Along these lines, the ISO/IEC Joint Technical Committee 1 subcommittee 37 on biometrics has just initiated projects on image quality assessment and face-aware capture.

Next steps: NIST expects to publish a first report on demographic dependencies in face recognition in 2019. This will include the effects of age, sex and race.

Technical Summary

▷ **Rank-based accuracy:** The inset table shows false negative "miss rates" realized when searching a 12 million person gallery populated with FRVT 2018 mugshots. The two most accurate algorithms fail to return the correct mate somewhere within the top 50 ranks in fewer than 0.1% of searches (Table 1, rows 1,2). This is achieved for galleries populated with multiple images per person. In the case where only the most recent image is present the miss rate is modestly higher (rows 3,4). The mates are almost always at rank 1, so in cases where only very short candidate lists must be used, the rank-1 miss rate is barely higher 0.12% (row 5) which again modestly rises when persons are enrolled with a single image (row 7). All the miss rates are measured over a fixed set of 154 549 searches, and the lowest false negative error rate recorded in this report (0.038%, row 10) corresponds to just 58 misses. Given such low error rates, what misses remain? By inspection they arise in five categories, those due to: a) ageing i.e. long-term time lapse between images; b) images of injured individuals e.g. bruised or bandaged faces; c) the presence of a second face e.g. printed on a T-shirt; d) images of some object that is not a face; e) profile-view images, and f) actual clerical ID label errors. As discussed in section 3.8.2, the first three categories are legitimately part of a test designed to measure accuracy on portrait images collected in law-enforcement settings. The latter three categories, however, should not be included in a test that is attempting to measure accuracy on only frontal images. Thus, by removing all known images in those categories, the rightmost column shows error rates that would be attainable in an application where exclusively frontal portrait images were collected without identity labeling errors.

	Investigation miss rate at	Num-subjects	Enrolled image	Num-images	Algorithm	FNIR Raw	FNIR Corrected
1	Rank-50	12M	Lifetime	26.1M	NEC-2	0.09%	0.09%
2	Rank-50	12M	Lifetime	26.1M	Microsoft-5	0.06%	**0.06%**
3	Rank-50	12M	Recent	12M	NEC-2	0.25%	**0.08%**
4	Rank-50	12M	Recent	12M	Microsoft-5	0.21%	0.09%
5	Rank-1	12M	Lifetime	26.1M	NEC-2	0.14%	**0.12%**
6	Rank-1	12M	Lifetime	26.1M	Microsoft-5	0.25%	0.24%
7	Rank-1	12M	Recent	12M	NEC-2	0.31%	**0.13%**
8	Rank-1	12M	Recent	12M	Microsoft-5	0.52%	0.37%
9	Rank-50	640K	Lifetime	1.25M	NEC-2	0.08%	0.08%
10	Rank-50	640K	Lifetime	1.25M	Microsoft-5	0.04%	**0.04%**

Table 1: Rank-based accuracy floor 2018.

Error rates today are two orders of magnitude below what they were in 2010, a massive reduction that stems from wholesale replacement of the old algorithms with those based on (deep) convolutional neural networks (CNNs). This constitutes a revolution rather than the evolution that defined the period 2010-2013. The rapid innovations around CNN architectures and loss functions including, both proprietary and published in the academic literature[1], may yet produce further gains. Even without that possibility, the results imply that prospective end-users should establish whether installed algorithms pre-date the development of the prototypes evaluated here and inquire with suppliers on availability of the latest versions. The gains mean that searches that had previously failed to yield candidates may now do so, such that unsolved cases could be revisited.

Given this impressive achievement - close to perfect recognition - an advocate might claim that frontal face recognition is a solved problem, a statement that should be refuted with the following context and caveats:

- ▷ **Algorithm accuracy spectrum:** Many algorithms do not achieve the low error rates tabulated above, and while many of those may still be useful and valuable to end-users, only the most accurate excel on poor quality images and those collected long after the initial enrollment sample.

- ▷ **Versioning:** While results for up to seven algorithms from each developer are reported here, the intra-provider accuracy variations are usually smaller than the inter-provider variations. That said different versions give order of magnitude fewer misses. Some developers demonstrate speed-accuracy tradeoffs[2]. See Figs. 17, 18.

[1] For example, Resnets [11], Inception [23], very deep networks [18, 21] and spatial transformers.

[2] NEC-0 prepares templates much faster than NEC-2 but gives twenty times more misses. Dermalog-5 executes a template search much more quickly than Dermalog-6 but is also much less accurate.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

- ▷ **Quality:** The low error rates here are attained using mostly excellent cooperative live-capture mugshot images collected with an attendant present. Recognition in other circumstances, particularly those without a dedicated photographic environment and human or automated quality control checks, will lead to declines in accuracy. This is documented here for poorer quality webcam images and unconstrained "wild" images.

- ▷ **Low similarity scores:** In thousands of cases the correct gallery image is returned at rank 1 but its similarity score is nevertheless low, below some operationally required score threshold. This does not matter when face recognition is used for "lead generation" in investigational applications because human reviewers are specifically required to review potentially long candidate lists and the threshold is effectively 0. In applications where search volumes are higher and labor is not available to review the results from searches, a higher threshold can be applied. This reduces the length of candidate lists and false positive identification rates at the expense of increased false negative miss rates. The tradeoff between the two error rates is reported extensively later.

- ▷ **Population size:** As the number of enrolled subjects grows, some mates are displaced from rank one, decreasing accuracy. As tabulated later for N up to 12 million, false negative rates generally rise slowly with population size.

- ▷ **Database integrity:** An operational error rate should be added to all false negative rates in this report reflecting the proportion of images in a real database that are un-matchable. Such anomalies arise from images that: do not contain a face; include multiple persons; cannot be decoded; are rotated by 90° or 180°; depict a face on clothing; and others introduced by a long tail of various clerical errors. While the mugshot trials in this report have been constructed to minimize such effects, they are a real problem in actual operations.

▷ **Threshold-based accuracy:** Recognition accuracy is very strongly dependent on the algorithm and, more generally, on the developer of the algorithm. False negative error rates in a particular scenario range from a few tenths of one percent to beyond fifty percent. This is tabulated exhaustively later: For example Table 22 shows accuracy across datasets. The inset figure here compares algorithms on mugshot searches in a consolidated gallery of 12 million subjects and 26.1 million photos. In positive or negative identification applications, a score threshold is set to limit the rate at which non-mate searches produce false positives. This has the consequence that some mated searches will report the mate below threshold, i.e. a miss, even if it is at rank 1. The utility of this is that many non-mated searches will usually not return any candidate identities at all. As the

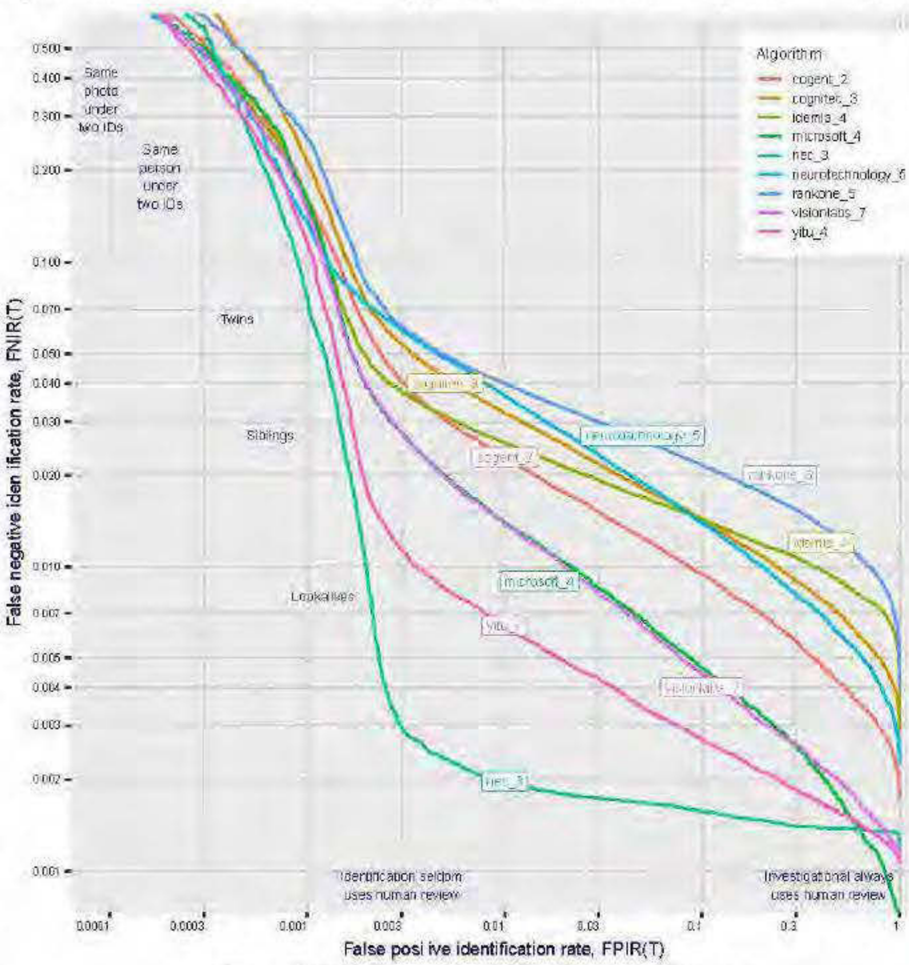


Figure 1: Miss rates across the false positive range

inset error-tradeoff characteristic
shows, investigational miss rates on the right side are very low but then rise steadily (in the center region) as threshold is increased to support "lights-out" applications, and ultimately rise quickly (left side) as discussed below. Thus, if we demand that just one in one thousand non-mate searches produce any false positives, the most accurate algorithm there (NEC-3) would fail on 7.9% of mated searches. Even though the graph shows results for the most accurate algorithms, all but two would fail to find the mate in more than 10% of mated searches. While the NEC algorithm produces a relatively flat error tradeoff until the threshold is raised to limit false positives to about 1 in 400 non-mated searches[3]

Thereafter, as the threshold is raised to further reduce false positives, miss rates rise rapidly. This means that low false positive identification rates are inaccesible with these algorithms, a result that does not apply for ten-finger identification algorithms. The rapid rise occurs because the lower mate scores are mixed with very high non-mate scores, the low scores from poor image quality and ageing, the high non-mates from the presence of lookalikes persons (doppelgangers), twins (discussed next) and, ultimately, the presence of a few unconsolidated subjects i.e. persons present under multiple IDs.

▷ **False positives from twins:** By enrolling 640 000 mugshots, adding photos of one twin, and then searching photos of those subjects and their twin the inset figure shows, for one typical algorithm, the similarity is generally greater when searching twins against themselves (A) than when searching twins against their sibling (B) but very often still above even stringent thresholds i.e. those corresponding to one in one thousand searches producing a false positive. Thus twins will very often produce a high-scoring non-match on a candidate list and a false alarm in an online identification system. The plot shows that some fraternal twins are correctly rejected at those thresholds - these are largely from different sex twins (at center). Figure 21 shows substantially similar behavior for all algorithms tested. In an investigative search, a twin would typically appear at rank 1, or rank 2 if their sibling happened to also be the gallery. Twins (and triplets etc.) constituted 3.3% of all live births [17] in recent years[4], and because that number is higher today than when the individuals in current adult databases were born, the false positives that arise from twins are now, and will increasingly be, an operational problem. Relative to the United States, twins are born with considerable regional variation. For example they are much less common in East Asia, and much more common in Sub-Saharan Africa [22]. The presence of twins in the mugshot database is inevitable given its size, around 12.3 million people. As this is not an insignificant sample of the domestic United States population, people with other familial ties will be present also. The data was collected over an extended period and because location information is not available, we are unable to estimate the proportion of the domestic population that is present in the dataset. However, if we assume twins are neither more or less disposed to arrest than the general population, we can estimate that hundreds of thousands of individuals in the dataset are twins. This will affect false positive rates because we randomly set aside 331 254 individuals for nonmate searches, and some proportion of those will be twins with siblings in the gallery.

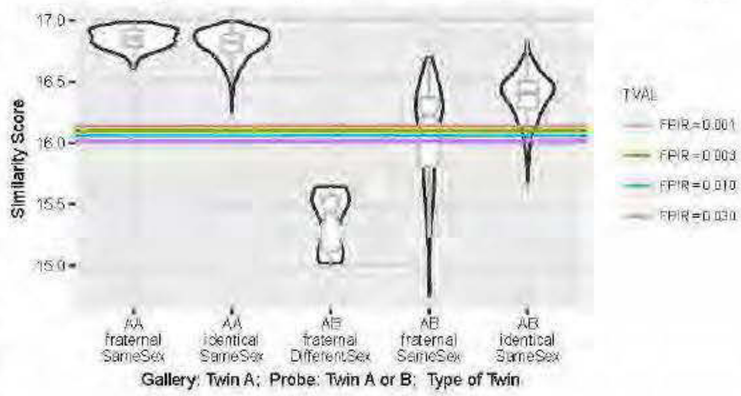


Figure 2: Intra- and inter-twin scores

[3] The gallery size here is 12 million people, 26.1 million images. Given 331 254 non-mated searches, an exhaustive implementation of one-too-many search would execute 8.6 trillion comparisons. At a false positive identification rate of 0.0025 the number of false positives is, to first order, 828 corresponding to single-comparison false match rate of 828 / 8.6 trillion = 9.6 10^{-11} i.e. about 1 in 10 billion. Strictly this FMRR computation meaningful only for algorithms that implement 1:N search using N 1:1 comparisons, which is not always the case.
[4] See the CDC's National Vital Statistics Report for 2017: https://www.cdc.gov/nchs/data/nvsr/nvsr67/nvsr67_08-508.pdf

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

▷ **False negatives from ageing:** A large source of error in long-run applications where subjects are not re-enrolled on a schedule is ageing. This is a function of the time elapsed between photographs. Change in facial appearance causes recognition similarity scores to decline such that over the longer term, accuracy will decline. All faces age and while this usually proceeds in a graceful and progressive manner, drug use can accelerate this [30]. Elective surgery may be effective in delaying it although this has not been formally quantified with face recognition. As ageing is essentially unavoidable, it can only be mitigated by scheduled re-capture, as in passport re-issuance. To quantify ageing effects, we used the more accurate algorithms to enroll the earliest image of 3.1 million adults and then search with 10.3 million newer photos taken up to 18 years after the the initial enrollment photo. In the inset table, accuracy is seen to degrade progressively with time, as mate scores decline and non-mates displace mates from rank 1 position. More accurate algorithms tend to be less sensitive to ageing. The more accurate algorithms give fewer errors after 18 years of ageing than middle tier algorithms give after four. Note also we do not quantify an ageing rate - more formal methods [2] borrowed from the longitudinal analysis literature have been published for doing so (given suitable repeated measures data). See Figures 62, 72 and 77.

Algorithm	Metric, FNIR@	(0,2]	(2,4]	(4,6]	(6,8]	(8,10]	(10,12]	(12,14]	(14,18]
nec-2	Rank = 1	0.3	0.4	0.4	0.4	0.4	0.5	0.6	0.4
microsoft-4	Rank = 1	0.3	0.5	0.6	0.7	0.9	1.0	1.3	1.6
yitu-4	Rank = 1	0.6	0.8	0.8	0.8	0.9	1.1	1.5	2.1
everai-3	Rank = 1	0.5	0.7	0.9	1.1	1.3	1.5	1.8	2.2
idemia-4	Rank = 1	1.1	1.5	1.9	2.3	2.8	3.1	3.7	5.1
cogent-3	Rank = 1	0.8	1.1	1.3	1.5	1.7	1.9	2.4	3.1
cognitec-2	Rank = 1	1.0	1.4	1.7	2.0	2.4	2.6	3.1	3.9
nec-2	FPIR = 0.001	0.7	0.9	1.1	1.3	1.5	1.7	2.1	2.7
microsoft-4	FPIR = 0.001	2.7	4.7	7.2	10.1	12.9	16.1	20.5	25.9
yitu-4	FPIR = 0.001	1.2	2.0	3.1	4.7	6.7	9.6	14.2	20.1
everai-3	FPIR = 0.001	3.5	6.2	9.3	12.9	16.2	19.6	24.1	29.2
idemia-4	FPIR = 0.001	3.7	5.9	8.3	11.0	13.4	15.8	19.1	24.8
cogent-3	FPIR = 0.001	5.8	9.7	14.2	19.2	23.8	28.4	34.4	42.1
cognitec-2	FPIR = 0.001	5.2	8.8	12.7	17.1	21.0	24.6	29.2	35.3

Table 2: Impact of ageing on accuracy.

▷ **Image quality matters:** Poor quality photographs undermine recognition, either because the imaging system is poor (lighting, camera, etc.) or because the subject mis-presents to the camera (head orientation, facial expression, occlusion, etc.). Imaging problems can be mitigated by design i.e. ensuring adherence to long-standing face image capture standards. Presentation problems, however, must be detected at capture time, either by the photographer, or by an automated system, and recapture performed. The most accurate algorithms in FRVT are highly tolerant of image quality problems. This derives from the invariances afforded by CNN-based algorithms, and this is the fundamental reason why accuracy has improved since 2013. For example, the Microsoft algorithms are can match many profile-view images to frontal mugshots - see Figures 100 and 102. As the inset table shows, rank-1 false negative identification rates are much higher with wild images than webcams and, in turn, mugshots. Further, even with the most capable algorithms, comparison scores are lower with unconstrained images, so that when (high) thresholds are necessary to limit false positives, here to 1 in 100 searches, error rates are very high. Such figures should guide prospective users of face recognition to consider whether face recognition can meet a formal written accuracy requirement.

Algorithm	Metric, FNIR@	Wild	Mugshot	Webcam
cognitec-3	Rank = 1	5.1	0.9	2.5
everai-3	Rank = 1	3.8	0.5	1.9
idemia-5	Rank = 1	4.4	1.1	3.9
microsoft-5	Rank = 1	3.3	0.3	1.1
nec-3	Rank = 1	8.8	0.3	1.0
ntechlab-6	Rank = 1	3.8	0.6	1.7
visionlabs-5	Rank = 1	4.3	0.4	1.9
yitu-4	Rank = 1	4.4	0.4	0.8
cognitec-3	FPIR = 0.01	32.5	2.8	10.0
everai-3	FPIR = 0.01	35.7	1.8	6.0
idemia-5	FPIR = 0.01	34.0	2.8	10.2
microsoft-5	FPIR = 0.01	34.4	1.2	4.1
nec-3	FPIR = 0.01	38.0	0.4	1.3
ntechlab-6	FPIR = 0.01	38.1	2.1	5.9
visionlabs-5	FPIR = 0.01	34.4	2.2	8.7
yitu-4	FPIR = 0.01	30.6	0.7	1.7

Table 3: Impact of image quality on accuracy.

▷ **Accuracy in large populations:** This report documents identification accuracy in galleries contaimining up to 12 million people and 26.1 million images. False negative rates climb very slowly as population size increases. For the most accurate algorithm, NEC-2, when searching a database of size 640 000, about 0.26% of searches fail to produce the correct mate as its best hypothesized identity. In a database of 12 000 000 this rises to just 0.31%. This benign growth in miss rates is fundamentally the reason for the utility of face recognition in large scale one-to-many search applications. See Table 12 and Figure 22.

The reason for this is that as more identities are enrolled into an database, the possibility of a false positive increases due to lookalike faces that yield extreme values from the right tail of the non-mate score distribution. However, these scores are lower than most mate scores such that when an identification algorithm is configured with a threshold of zero (so human adjudication is always necessary), rank-one identification miss rates scale very favorably with population size, N, growing slowly, approximately as a power law, aN^b with $b \ll 1$. This dependency was first noted in 2010. Depending on the algorithm, the exponent b for mugshot searches is low, around 0.06 for the some of the more accurate algorithms with up to 12 million identities. See Table 12.

In any case, variations in accuracy with increasing population size are small relative to both ageing and algorithm choice. See Figure 20.

▷ **Utility of adjudicating long candidate lists:** In the regime where a system is configured with a threshold of zero, and where human adjudication is always necessary, the reviewer will find some mates quite far down candidate lists. This usually occurs because either the probe image or its corresponding enrolled mate image have poor quality, or large time-lapse. The accuracy benefits of traversing say 50 candidates versus just the first one is broadly a reduction in error by up to a factor of two. See Figure 30 and compare Tables 12 and 13.

However, accuracy from the leading algorithm is now so high - mates that in 2013 were placed at rank > 1, are now at rank 1 - such that reviewers can expect to review substantially fewer candidates. Note, however, for the proportion of searches where there is no mate, reviewers might still examine all candidates, fruitlessly. This report does not address the issue of human error in adjudicating candidates produced in one-to-many searches.

▷ **Utility of enrolling multiple images per subject:** We run three kinds of enrollment: First, by enrolling just the most recent image; second by creating a single template from a person's full lifetime history of images; and third by enrolling multiple images of a person separately, as though under different identities. The overall effect is that the enrollment of multiple images yields as much as a factor of two lower miss rates. This occurs due to higher information content and because the most recent image may sometimes be of poorer quality than historical images. See Table 12.

Gains depend on the number of available images: FNIR drops steadily. Some algorithms reduce FPIR or maintain it - the desirable behaviors - but others give higher false positive rates. See Figures leading up to Figure 87.

▷ **Reduced template sizes:** There has been a trend toward reduced template sizes, i.e. a smaller feature representation of an image. In 2014, the most accurate algorithm used a template of size 2.5KB; the figure in 2018 is around 1600 bytes. Close competitors produce templates of size 256, 364, 512, and about 2KB bytes. In 2014, the leading competitors had templates of size 4KB to 8KB. Some algorithms, when enrolling more than one image of a person, produce a template whose size is independent of the number of images given to the algorithm. This can be achieved by selecting a "best" image, or by integrating (fusing) information from the images. See Table 16.

▷ **Template generation times:** Template generation times, as measured on a single circa-2016 server processor core [5], vary from below 20 milliseconds up to nearly 1 second. This wide variation across developers may be relevant to end-users who have high-volume workflows. There has not been a wide downward trend since 2014. Note that speed may be expedited over the figure reported here by exploiting new vector instructions on recent chips. Note that GPUs were not used and, while indispensable for training CNNs, are not necessary for feeding an image forward through a network. See Table 16.

▷ **Search duration and scalability:** Template search times, as measured on circa-2016 Intel server processor cores,

[5] Intel Xeon CPU E5-2630 v4 running at 2.20GHz.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8238

vary massively across the industry. For a database of size 1 million subjects, and the more accurate implementations, durations range from below 1 to 500 milliseconds, with other less accurate algorithms going much slower still. Several algorithms exhibit sublinear search time i.e. the duration does not double with a doubling of the enrolled population size, N. This was noted also in 2014. This has improved in 2018, however, such that close-to-logarithmic growth is evident for several developers' algorithms and extremely fast search. The consequence of this is that as N increases even the fastest linear algorithm (NEC-3) will quickly become much slower than the strongly sublinear algorithms. For the Dermalog-5 algorithm, search of a template against a database of N = 12 million images takes 850 microseconds on a single core of a contemporary CPU. That number is faster than any other algorithm even with the smallest gallery we tested (N = 640 000). See Table 6 and Figure 111.

▷ **Accuracy gains June - October 2018** NIST Interagency Report 8238 documented massive gains from 2013 to 2018. This report shows most developers achieved gains over the four month interval between June and October 2018. For a set of 12 million subjects enrolled with their most recent mugshot image, the inset table shows, for selected algorithms, the proportion of searches where mates are not returned against the given criteria (column 2). The result is that substantial reductions in false negatives - by a factor of two or more - were realized by algorithms submitted by Cogent, Cognitec, Dermalog, Hikvision, Innovatrics, NEC, Rank One, Shaman, Tiger-IT, and Vigilant Solutions. In particular, in this same time period one developer, NEC, which had produced broadly the most accurate algorithms in 2010 and 2013, submitted algorithms that are substantially more accurate than their June 2018 versions, and on many measures are now the most accurate. A number of other developers produced slightly less accurate implementations.

Application	Metric	Algorithm		FNIR
Mode: Mugshot	Miss rate	Date	Name	
Investigation	at Rank=1	2018-JUN	NEC-0	3.20%
Investigation	at Rank=1	2018-OCT	NEC-2	0.31%
Investigation	at Rank=1	2018-JUN	Microsoft-4	0.45%
Investigation	at Rank=1	2018-OCT	Microsoft-5	0.52%
Investigation	at Rank=1	2018-JUN	Yitu-2	0.55%
Investigation	at Rank=1	2018-OCT	Yitu-5	0.55%
Identification	at FPIR=0.001	2018-JUN	NEC-0	20.0%
Identification	at FPIR=0.001	2018-OCT	NEC-3	5.8%
Identification	at FPIR=0.001	2018-JUN	Microsoft-4	15.8%
Identification	at FPIR=0.001	2018-OCT	Microsoft-6	15.6%
Identification	at FPIR=0.001	2018-JUN	Yitu-2	12.4%
Identification	at FPIR=0.001	2018-OCT	Yitu-5	11.1%

Table 4: Accuracy gains since June - October 2018

See Tables 16 and 19, and Figure 19.

▷ **Non-technical considerations:** Recognition accuracy is likely the most important technical indicator for an algorithm. But even among the more accurate developers accuracy, template size, and resource consumption vary widely. This, incidentally, implies that technological diversity remains, that there is no consensus on approach and that algorithms are not commoditized. But beyond the performance statements in this report, face recognition outcomes in complete systems will be influenced by things like code and model size, software maturity, extensibility, reliability, ease of integration and maintenance, cost, availability of monitoring tools, and support for human review of true and false matches using, for example, capable graphical user interfaces.

▷ **Conclusions:** As with other biometrics, accuracy of facial recognition implementations varies greatly across the industry. Absent other performance or economic parameters, users should prefer the most accurate algorithm. Note that accuracy, and algorithm rankings, vary somewhat with the kinds of images used and the mode of operation: investigation with zero threshold; or identification with high threshold.

▷ **Supplementary Data:** This document is accompanied by a supplement that includes a three page report for each of the algorithms evaluated. Each report includes various performance plots pertinent to the particular algorithm under test. The supplement, which currently runs to more than 600 pages, is available from the same webpage as this report.

Release Notes

FRVT Activities: NIST restarted FRVT's one-to-many track in February 2018, inviting participants to send up to seven prototype algorithms. Since February 2017, NIST has been evaluating one-to-one verification algorithms on an ongoing basis. This allows developers to submit updated algorithms to NIST at any time but no more frequently than four calendar months. This more closely aligns development and evaluation schedules. Results are posted to the web within a few weeks of submission. Details and full report are linked from the Ongoing FRVT site.

FRVT Reports: The results of the FRVT appear in the series NIST Interagency Reports tabulated below. The reports were developed separately and released on different schedules. In prior years NIST has mostly reported FRVT results as a single report; this had the disadvantage that results from completed sub-studies were not published until all other studies were complete.

Date	Link	Title	No.
2014-03-20	PDF	FRVT Performance of Automated Age Estimation Algorithms	7995
2015-04-20	PDF	Face Recognition Vendor Test (FRVT) Performance of Automated Gender Classification Algorithms	8052
2014-05-21	PDF	FRVT Performance of face identification algorithms	8009
2017-03-07	PDF	Face In Video Evaluation (FIVE) Face Recognition of Non-Cooperative Subjects	8173
2017-11-23	PDF	The 2017 IARPA Face Recognition Prize Challenge (FRPC)	8197
2018-04-13	WWW	Ongoing Face Recognition Vendor Test (FRVT)	Draft

Details appear on pages linked from `https://www.nist.gov/programs-projects/face-projects`.

Appendices: This report is accompanied by appendices which present exhaustive results on a per-algorithm basis. These are machine-generated and are included because the authors believe that visualization of such data is broadly informative and vital to understanding the context of the report.

Typesetting: Virtually all of the tabulated content in this report was produced automatically. This involved the use of scripting tools to generate directly type-settable LaTeX content. This improves timeliness, flexibility, maintainability, and reduces transcription errors.

Graphics: Many of the Figures in this report were produced using the ggplot2 package running under R, the capabilities of which extend beyond those evident in this document.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

Contents

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

1 Introduction

One-to-many identification represents the largest market for face recognition technology. Algorithms are used across the world in a diverse range of biometric applications: detection of duplicates in databases, detection of fraudulent applications for credentials such as passports and driving licenses, token-less access control, surveillance, social media tagging, lookalike discovery, criminal investigation, and forensic clustering.

This report contains a breadth of performance measurements relevant to many applications. Performance here refers to accuracy and resource consumption. In most applications, the core accuracy of a facial recognition algorithm is the most important performance variable. Resource consumption will be important also as it drives the amount of hardware, power, and cooling necessary to accommodate high volume workflows. Algorithms consume processing time, they require computer memory, and their static template data requires storage space. This report documents these variables.

1.1 Open-set searches

FRVT tested open-set identification algorithms. Real-world applications are almost always "open-set", meaning that some searches have an enrolled mate, but some do not. For example, some subjects have truly not been issued a visa or drivers license before; some law enforcement searches are from first-time arrestees[6]. In an "open-set" application, algorithms make no prior assumption about whether or not to return a high-scoring result, and for a mated search, the ideal behaviour is that the search produces the correct mate at high score and first rank. For a non-mate search, the ideal behavior is that the search produces zero high-scoring candidates.

Too many academic benchmarks execute only closed-set searches. The proportion of mates found in the rank one position is the default accuracy metric. This hit rate metric ignores the score with which a mate is found; weak hits count as much as strong hits. This ignores the real-world imperative that in many applications it is necessary to elevate a threshold to reduce the number of false positives.

2 Evaluation datasets

This report documents accuracy for four kinds of images - mugshots, webcam, profiles and wild - as described in the following sections.

2.1 Mugshot images

The main mugshot dataset used is referred to as the FRVT 2018 set. This set was collected over the period 2002 to 2017 in routine United States law enforcement operations. This set has been extracted from a larger operational parent set by excluding non-face images, and setting aside webcam and profile-view images, for use in separate tests.

NIST Interagency Report 8238 includes a comparison of this set of mugshots with the smaller and easier sets of mugshots used in tests run in 2010 and 2014.

[6]Operationally closed-set applications are rare because it is usually not the case that all searches have an enrolled mate. One counter-example, however, is a cruise ship in which all passengers are enrolled and all searches should produce exactly one identity. Another example is forensic identification of dental records from an aircraft crash.

▷ **Mugshots**: Mugshots comprise about 86% of the database. They have reasonable compliance with the ANSI/NIST ITL1-2011 Type 10 standard's subject acquisition profiles levels 10-20 for frontal images [28]. The most common departure from the standard's requirements is the presence of mild pose variations around frontal - the images of Figure 3 are typical. The images vary in size, with many being 480x600 pixels with JPEG compression applied to produce filesizes of between 18 and 36KB with many images outside this range, implying that about 0.5 bits are being encoded per pixel.

▷ **Profile images**: Profile-view images have been collected in law enforcement for more than 100 years, as human capability is improved with orthogonal information. The profile images used in this report were collected during the same session as the frontal mugshot photograph, in the same standardized photographic setup. These would not therefore be used with automated face recognition. A small subset, 200 000 images, were set aside for testing.

▷ **Webcam images**: The remaining 14% of the images were collected using an inexpensive webcam attached to a flexible operator-directed mount. These images are all of size 240x240 pixels, that are in considerable violation of most quality-related clauses of all face recognition standards. As evident in the figure, the most common defects are non-frontal pose (associated with the rotational degrees of freedom of the camera mount), low contrast (due to varying and intense background lights), and poor spatial resolution (due to inexpensive camera optics) - see examples in Fig 4. The images are overly JPEG compressed, to between 4 and 7KB, implying that only 0.5 to 1 bits are being encoded per color pixel.

Example images are shown in Figures 3, 4 and 5 These are drawn from NIST Special Database 32 which may be downloaded here.

These images were partitioned in galleries and probesets for the various experiment listed in Table 5.

Figure 3: Six mated mugshot pairs representative of the FRVT-2014 (LEO) and FRVT-2018 datasets. The images are collected live, i.e. not scanned from paper. Image source: NIST Special Database 32

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

Figure 5: **[Profile views]** *The three images are a frontal enrollment, subsequent frontal probe, and same-session ninety degree profile view. While collection of both frontal and profile views has been typical in law enforcement for more than a century, the recognition of profile to frontal views has essentially been impossible. However, reasonbly high accuracy results is now possible - see section E.*

Figure 4: Twelve webcam images representative of probes against the FRVT-2018 mugshot gallery. The first eight images are four mated pairs. Such images present challenges to recognition including pose, non-uniform illumination, low contrast, compression, cropping, and low spatial sampling rate. Image source: NIST Special Database 32

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

2.2 Unconstrained "wild" images

In addition to portrait-styled mugshots, algorithms were also evaluated on a "wild" dataset composed of non-cooperative and unconstrained photojournalism and amateur photography imagery. The images are closely cropped from the parent images as shown in Figure 6. A portion of the images are collected by professional photographers and as such are captured, and selected, to not exhibit exposure and focus problems. Some of the photos were downloaded from websites with substantial amateur photographer imagery, which may contain images that do exhibit exposure and focus problems. Resolution varies widely as these images were downloaded from the internet with varying resampling and compression practices. The primary difficulties for face recognition is unconstrained yaw and pitch pose variation, with some images extending to profile view. Additionally faces can be occluded, including by hair and hands.

The images are cropped prior to passing them to the algorithm. The cropping is done per human-annotated rectangular bounding boxes. The algorithm must further localize the face and extract features. In many cases, there were multiple images of the subject provided to the algorithm, and the output was a single template representation of the subject.

$N_P = 332\,574$ subjects were searched against two galleries, where the number of enrolled subjects in each gallery were $N_{G1} = 1\,106\,777$ and $N_{G2} = 1\,107\,778$. Both gallery and search images were composed of unconstrained wild imagery.

Figure 6: Examples of "in the wild" stills. The top row gives the full original images; the second row gives the manually specified face region that is cropped and passed to the algorithms. The source images in this figure are attributed to, from left, Rita Molnr, Eva Rinaldi, and Gage Skidmore under the [cc-by-sa-2.5], [cc-by-sa-2.0], [cc-by-sa-3.0] creative commons licenses respectively.

2.3 Enrollment strategies

Many operational applications include collection and enrollment of biometric data from subjects on more than one occasion. This might be done on a regular basis, as might occur in credential (re-)issuance, or irregularly, as might happen in a criminal recidivist situation [4]. The number of images per person will depend on the application area. In civil identity credentialing (e.g. passports, driver's licenses), the images will be acquired approximately uniformly over time (e.g. ten years for a passport). While the distribution of dates for such images of a person might be assumed uniform, a number of factors might undermine this assumption[7]. In criminal applications, the number of images would depend on the number of arrests. The distribution of dates for arrest records for a person (i.e. the recidivism distribution) has been modeled using the exponential distribution but is recognized to be more complicated[8].

In any case, the 2010 NIST evaluation of face recognition showed that considerable accuracy benefits accrue with retention and use of *all* historical images [6].

To this end, the FRVT API document provides $K \geq 1$ images of an individual to the enrollment software. The software is tasked with producing a single proprietary undocumented "black-box" template[9] from the K images. This affords the algorithm an ability to generate a *model* of the individual, rather than to simply extract features from each image on a sequential basis.

As depicted in Figure 7, the i-th individual in the FRVT 2018 dataset has K_i images. These are labelled as x_k for $k = 1 \ldots K_i$ in chronological order of capture date. To measure the utility of having multiple enrollment images, this report evaluates three kinds of enrollment:

[7] For example, a person might skip applying for a passport for one cycle, letting it expire. In addition, a person might submit identical images (from the same photography session) to consecutive passport applications at five year intervals.
[8] A number of distributions have been considered to model recidivism, see for example [3].
[9] There are no formal face template standards. Template standards only exist for fingerprint minutiae - see ISO/IEC 19794-2:2011.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

Image				
Encounter	1	...	$K_i - 1$	K_i
Capture Time	T_1	...	T_{K_i-1}	T_{K_i}
Role RECENT	Not used	Not used	Enrolled	Search
Role LIFETIME	Enrolled	Enrolled	Enrolled	Search

Figure 7: Depiction of the "recent" and "lifetime" enrollment types. Image source: NIST Special Database 32

- ▷ **Recent**: Only the second most recent image, x_{K_i-1} is enrolled. This strategy of enrollment mimics the operational policy of retaining the imagery from the most recent encounter. This might be done operationally to ameliorate the effects of face ageing. Obviously retaining only the most recent image should only be done if the identity of the person is trusted to be correct. For example, in an access control situation retention of the most recent successful *authentication* image would be hazardous if it could be a false positive.

- ▷ **Lifetime-consolidated**: All but the most recent image are enrolled, $x_1 \ldots x_{K_i-1}$. This subject-centric strategy might be adopted if quality variations exist where an older image might be more suitable for matching, despite the ageing effect.

- ▷ **Lifetime-unconsolidated**: Again all but the most recent image are enrolled $x_1 \ldots x_{K_i-1}$ but now separately, with different identifiers, such that the algorithm is not aware that the images are from the same face. This kind of event- or encounter-centric enrollment is very common when operational constraints preclude reliable consolidation of the historical encounters into a single identity. This aspect also prevents the recognition algorithm from a) building a holistic model of identity (as is common in speaker recognition systems) and b) implementing fusion, for example template-level fusion of feature vectors, or post-search score-level fusion. The result is that searches will typically yield more than one image of a person in the top ranks. This has consequences for appropriate metrics, as detailed in section 3.2.1

  NIST first evaluated this kind of enrollment in mid 2018, and the results tables include some comparison of accuracy available from all three enrollment styles.

In all cases, the most recent image, x_{K_i}, is reserved as the search image. For the 1.6 million subject enrollment partition of the FRVT 2018 data, $1 \leq K_i \leq 33$ with $K_i = 1$ in 80.1% of the individuals, $K_i = 2$ in 13.4%, $K_i = 3$ in 3.7%, $K_i = 4$ in 1.4%, $K_i = 5$ in 0.6%, $K_i = 6$ in 0.3%, and $K_i > 6$ is 0.2% for everyone else. This distribution is substantially dependent on United States recidivism rates.

We did not evaluate the case of retaining only the highest quality image, since automated quality assessment is out of scope for this report. We do not anticipate that such strategies will prove beneficial when the quality assessment apparatus is imperfect and unvalidated.

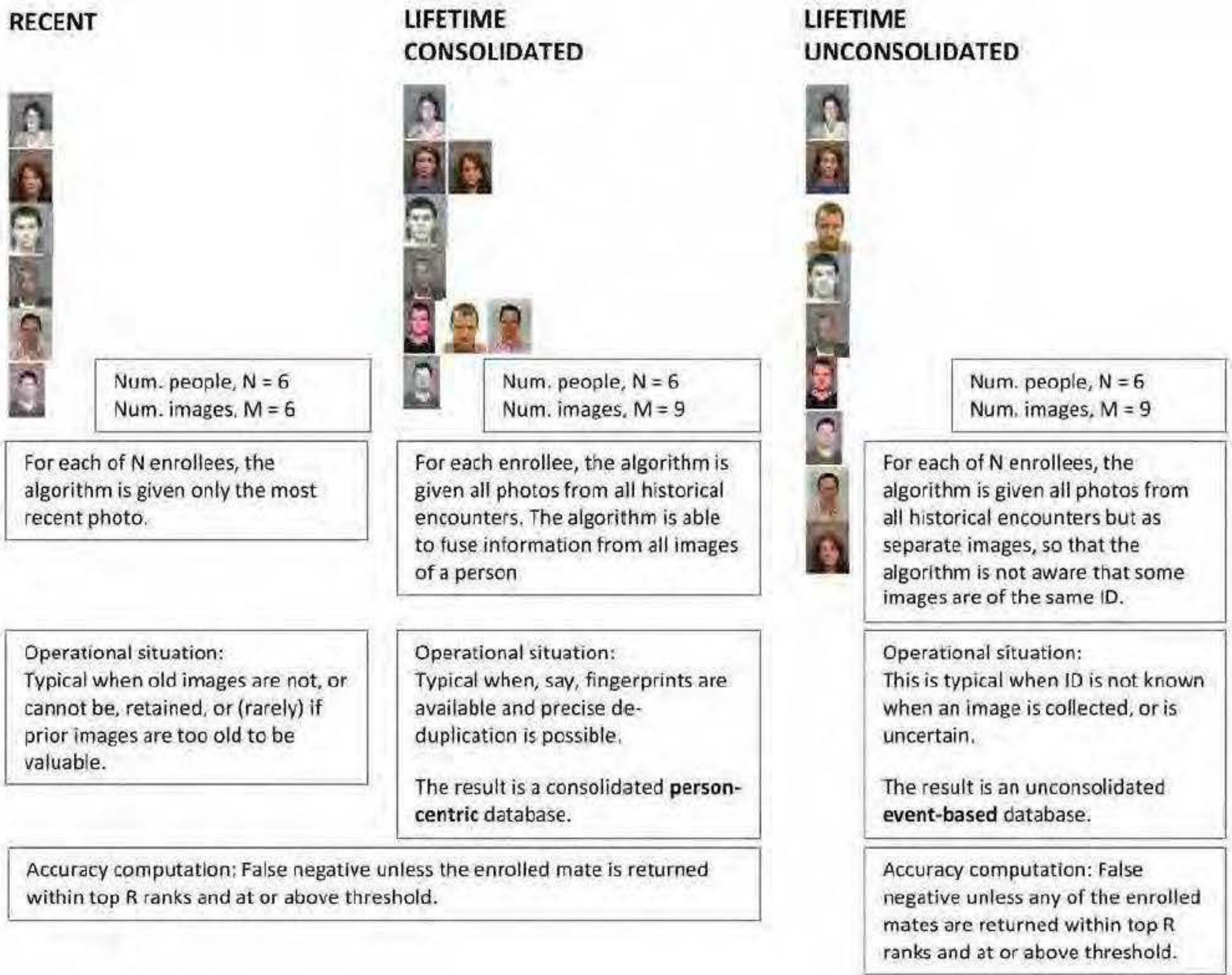


Figure 8: **Enrollment strategies**. *The figure shows the three kinds of enrollment databases examined in this report. Image source: NIST Special Database 32*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

	ENROLLMENT				SEARCH			
	TYPE SEE SECTION 2.3	POPULATION FILTER			MATE		NON-MATE	
			N-SUBJECTS	N-IMAGES	N-SUBJECTS	N-IMAGES	N-SUBJECTS	N-IMAGES
Mugshot trials from enrollment of single images								
1	RECENT	NATURAL	640 000	640 000	154 549	154 549	331 254	331 254
2	RECENT	NATURAL	1 600 000	1 600 000				
3	RECENT	NATURAL	3 000 000	3 000 000				
4	RECENT	NATURAL	6 000 000	6 000 000				
5	RECENT	NATURAL	12 000 000	12 000 000				
Mugshot trials from enrollment of lifetime images								
6	CONSOL	NATURAL	640 000	1 247 331				
7	CONSOL	NATURAL	1 600 000	3 351 206				
8	CONSOL	NATURAL	3 000 000	6 417 057				
9	CONSOL	NATURAL	6 000 000	12 976 185				
10	CONSOL	NATURAL	12 000 000	26 107 917				
11	UN-CONSOL	NATURAL	640 000	1 247 331				
12	UN-CONSOL	NATURAL	1 600 000	3 351 206				
Cross-domain								
13	MUGSHOTS AS ON ROW 2				82 106 WEBCAM	82 106 WEBCAM	331 254 WEBCAM	331 254 WEBCAM
Cross-view								
14	MUGSHOTS AS ON ROW 2				100 000 PROFILE	100 000 PROFILE	100 000 PROFILE	100 000 PROFILE
Ageing								
17	OLDEST	NATURAL	3 068 801	3 068 801	2 853 221	10 951 064	0	0

Table 5: **Enrollment and search sets.** *Each row summarizes one identification trial. Unless stated otherwise, all entries refer to mugshot images. The term "natural" means that subjects were selected without heed to demographics, i.e. in the distribution native to this dataset. The probe images were collected in a different calendar year to the enrollment image. Missing values in rows 2-12 are the same as in row 1.*

3 Performance metrics

This section gives specific definitions for accuracy and timing metrics. Tests of open-set biometric algorithms must quantify frequency of two error conditions:

- ▷ **False positives**: Type I errors occur when search data from a person who has never been seen before is incorrectly associated with one or more enrollees' data.
- ▷ **Misses**: Type II errors arise when a search of an enrolled person's biometric does not return the correct identity.

Many practitioners prefer to talk about "hit rates" instead of "miss rates" - the first is simply one minus the other as detailed below. Sections 3.1 and 3.2 define metrics for the Type I and Type II performance variables.

Additionally, because recognition algorithms sometimes fail to produce a template from an image, or fail to execute a one-to-many search, the occurrence of such events must be recorded. Further because algorithms might elect to not produce a template from, for example, a poor quality image, these failure rates must be combined with the recognition error rates to support algorithm comparison. This is addressed in section 3.5.

Finally, section 3.7 discusses measurement of computation duration, and section 3.8 addresses the uncertainty associated with various measurements. Template size measurement is included with the results.

3.1 Quantifying false positives

It is typical for a search to be conducted into an enrolled population of N identities, and for the algorithm to be configured to return the closest L candidate identities. These candidates are ranked by their score, in descending order, with all scores required to be greater than or equal to zero. A human analyst might examine either all L candidates, or just the top $R \leq L$ identities, or only those with score greater than threshold, T. The workload associated with such examination is discussed later, in 3.6.

False alarm performance is quantified in two related ways. These express how many searches produces false positives, and then, how many false positives are produced in a search.

False positive identification rate: The first quantity, FPIR, is the proportion of non-mate searches that produce an adverse outcome:

$$\mathrm{FPIR}(N,T) = \frac{\text{Num. non-mate searches where one or more enrolled candidates are returned with score at or above threshold}}{\text{Num. non-mate searches attempted.}} \tag{1}$$

Under this definition, FPIR can be computed from the highest non-mate candidate produced in a search - it is not necessary to consider candidates at rank 2 and above. FPIR is the primary measure of Type I errors in this report.

Selectivity: However, note that in any given search, several non-mate may be returned above threshold. In order to quantify such events, a second quantity, selectivity (SEL), is defined as the *number* of non-mates returned on a candidate list, averaged over all searches.

$$\mathrm{SEL}(N,T) = \frac{\text{Num. non-mate enrolled candidates returned with score at or above threshold}}{\text{Num. non-mate searches attempted.}} \tag{2}$$

where $0 \leq$ SEL(N, T) $\leq$L. Both of these metrics are useful operationally. FPIR is useful for targeting how often an adverse false positive outcome can occur, while SEL as a number is related to workload associated with adjudicating candidate lists. The relationship between the two quantities is complicated - it depends on whether an algorithm concentrates the false alarms in the results of a few searches or whether it disburses them across many. This was detailed in FRVT 2014, NISTIR 8009. It has not yet been detailed in FRVT 2018.

3.2 Quantifying hits and misses

If L candidates are returned in a search, a shorter candidate list can be prepared by taking the top $R \leq L$ candidates for which the score is above some threshold, $T \geq 0$. This reduction of the candidate list is done because thresholds may be applied, and only short lists might be reviewed (according to policy or labor availability, for example). It is useful then to state accuracy in terms of R and T, so we define a "miss rate" with the general name **false negative identification rate** (FNIR), as follows:

$$\mathrm{FNIR}(N,R,T) = \frac{\text{Num. mate searches with enrolled mate found outside top R ranks or score below threshold}}{\text{Num. mate searches attempted.}} \tag{3}$$

This formulation is simple for evaluation in that it does not distinguish between causes of misses. Thus a mate that is not reported on a candidate list is treated the same as a miss arising from face finding failure, algorithm intolerance of poor quality, or software crashes. Thus if the algorithm fails to produce a candidate list, either because the search

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

failed, or because a search template was not made, the result is regarded as a miss, adding to FNIR.

Hit rates, and true positive identification rates: While FNIR states the "miss rate" as how often the correct candidate is either not above threshold or not at good rank, many communities prefer to talk of "hit rates". This is simply the **true positive identification rate**(TPIR) which is the complement of FNIR giving a positive statement of how often mated searches are successful:

$$\mathrm{TPIR}(N, R, T) = 1 - \mathrm{FNIR}(N, R, T) \tag{4}$$

This report does not report true positive "hit" rates, preferring false negative miss rates for two reasons. First, costs rise linearly with error rates. For example, if we double FNIR in an access control system, then we double user inconvenience and delay. If we express that as decrease of TPIR from, say 98.5% to 97%, then we mentally have to invert the scale to see a doubling in costs. More subtly, readers don't perceive differences in numbers near 100% well, becoming inured to the "high nineties" effect where numbers close to 100 are perceived indifferently.

Reliability is a corresponding term, typically being identical to TPIR, and often cited in automated (fingerprint) identification system (AFIS) evaluations.

An important special case is the **cumulative match characteristic**(CMC) which summarizes accuracy of mated-searches only. It ignores similarity scores by relaxing the threshold requirement, and just reports the fraction of mated searches returning the mate at rank R or better.

$$\mathrm{CMC}(N, R) = 1 - \mathrm{FNIR}(N, R, 0) \tag{5}$$

We primarily cite the complement of this quantity, $\mathrm{FNIR}(N, R, 0)$, the fraction of mates *not* in the top R ranks.

The **rank one hit rate** is the fraction of mated searches yielding the correct candidate at best rank, i.e. CMC(N, 1). While this quantity is the most common summary indicator of an algorithm's efficacy, it is not dependent on similarity scores, so it does not distinguish between strong (high scoring) and weak hits. It also ignores that an adjudicating reviewer is often willing to look at many candidates.

3.2.1 False negative rates for unconsolidated galleries

As detailed in section 2.3 a common type of gallery, here referred to as the lifetime unconsolidate type, is populated with all images of an individual without any association between them. That is, the gallery construction algorithm is not provided with any ID labels that would support processing of a person's images jointly. This constrasts with the lifetime consolidate type where an algorithm may explicitly fuse features from multiple images of a person, or select a best image. In such cases, where the number of enrolled images is a random variable, we define two false negative rates as follows.

The first demands that the algorithm place any of the K_i mates in the top $R \geq 1$ ranks. The proportion of searches for which this does not occur forms a false negative identification rate:

$$\mathrm{FNIR}_{\mathrm{any}}(N, R, T) = 1 - \frac{\text{Num. mate searches where any enrolled mate is found in the top R ranks and at-or-above threshold}}{\text{Num. mate searches attempted.}} \tag{6}$$

The second demands that the algorithm place all K_i mates in the top $R \geq K_i$ ranks. The proportion of searches for

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

which this does not occur forms a false negative identification rate:

$$\mathrm{FNIR}_{\mathrm{all}}(N, R, T) = 1 - \frac{\text{Num. mate searches where all enrolled mates are found in the top R ranks and at-or-above threshold}}{\text{Num. mate searches attempted.}} \tag{7}$$

Placing all mates in the top ranks is a more difficult task than correctly retrieving any image, so it holds that: $\mathrm{FNIR}_{\mathrm{all}} \geq \mathrm{FNIR}_{\mathrm{any}}$. This is evident in the results presented for November 2018 algorithms in Tables starting at 25.

The information retrieval community might prefer to compute and plot *precision* and *recall*; this is a valid approach, but we advance the two metrics above because they relate to our normal definition of consolidated FNIR, and they cover the two extreme use-cases of wanting any hit vs. all hits.

3.3 DET interpretation

In biometrics, a false negative occurs when an algorithm fails to match two samples of one person a Type II error. Correspondingly, a false positive occurs when samples from two persons are improperly associated a Type I error.

Matches are declared by a biometric system when the native comparison score from the recognition algorithm meets some threshold. Comparison scores can be either similarity scores, in which case higher values indicate that the samples are more likely to come from the same person, or dissimilarity scores, in which case higher values indicate different people. Similarity scores are traditionally computed by fingerprint and face recognition algorithms, while dissimilarities are used in iris recognition. In some cases, the dissimilarity score is a distance possessing metric properties. In any case, scores can be either mate scores, coming from a comparison of one persons samples, or nonmate scores, coming from comparison of different persons samples.

The words "genuine" or "authentic" are synonyms for mate, and the word "impostor" is used as a synonym for nonmate. The words "mate" and "nonmate" are traditionally used in identification applications (such as law enforcement search, or background checks) while genuine and impostor are used in verification applications (such as access control).

An error tradeoff characteristic represents the tradeoff between Type II and Type I classification errors. For identification this plots false negative vs. false positive identification rates i.e. FNIR vs. FPIR parametrically with T. Such plots are often called detection error tradeoff (DET) characteristics or receiver operating characteristic (ROC). These serve the same function – to show error tradeoff – but differ, for example, in plotting the complement of an error rate (e.g. $\mathrm{TPIR} = 1 - \mathrm{FNIR}$) and in transforming the axes, most commonly using logarithms, to show multiple decades of FPIR. More rarely, the function might be the inverse of the Gaussian cumulative distribution function.

The slides of Figures 9 through 15 discuss presentation and interpretation of DETs used in this document for reporting face identification accuracy. Further detail is provided in formal biometrics testing standards, see the various parts of ISO/IEC 19795 Biometrics Testing and Reporting. More terms, including and beyond those to do with accuracy, appear in ISO/IEC 2382-37 Information technology – Vocabulary – Part 37: Harmonized biometric vocabulary.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

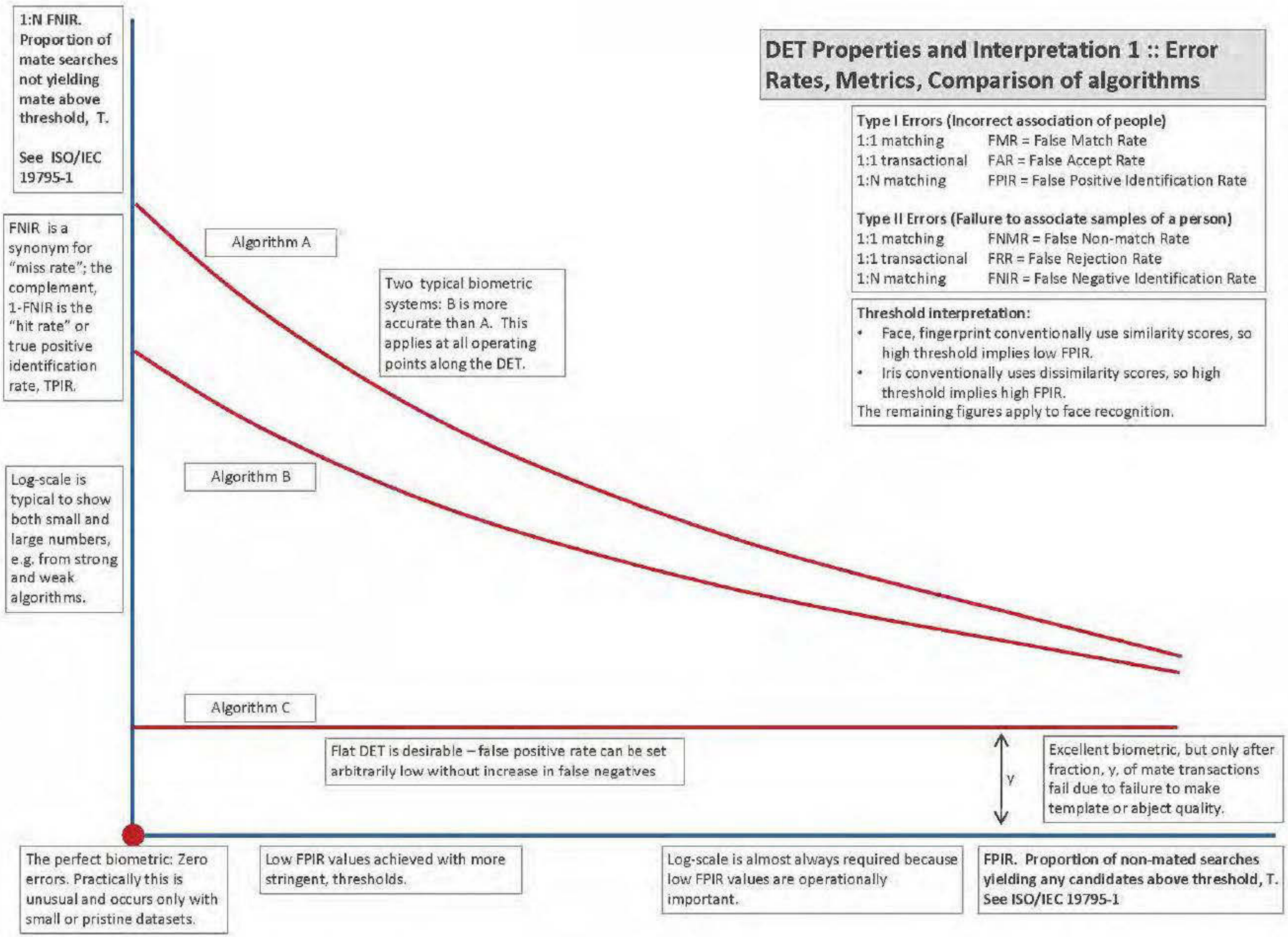


Figure 9: **DET as the primary performance reporting mechanism.**

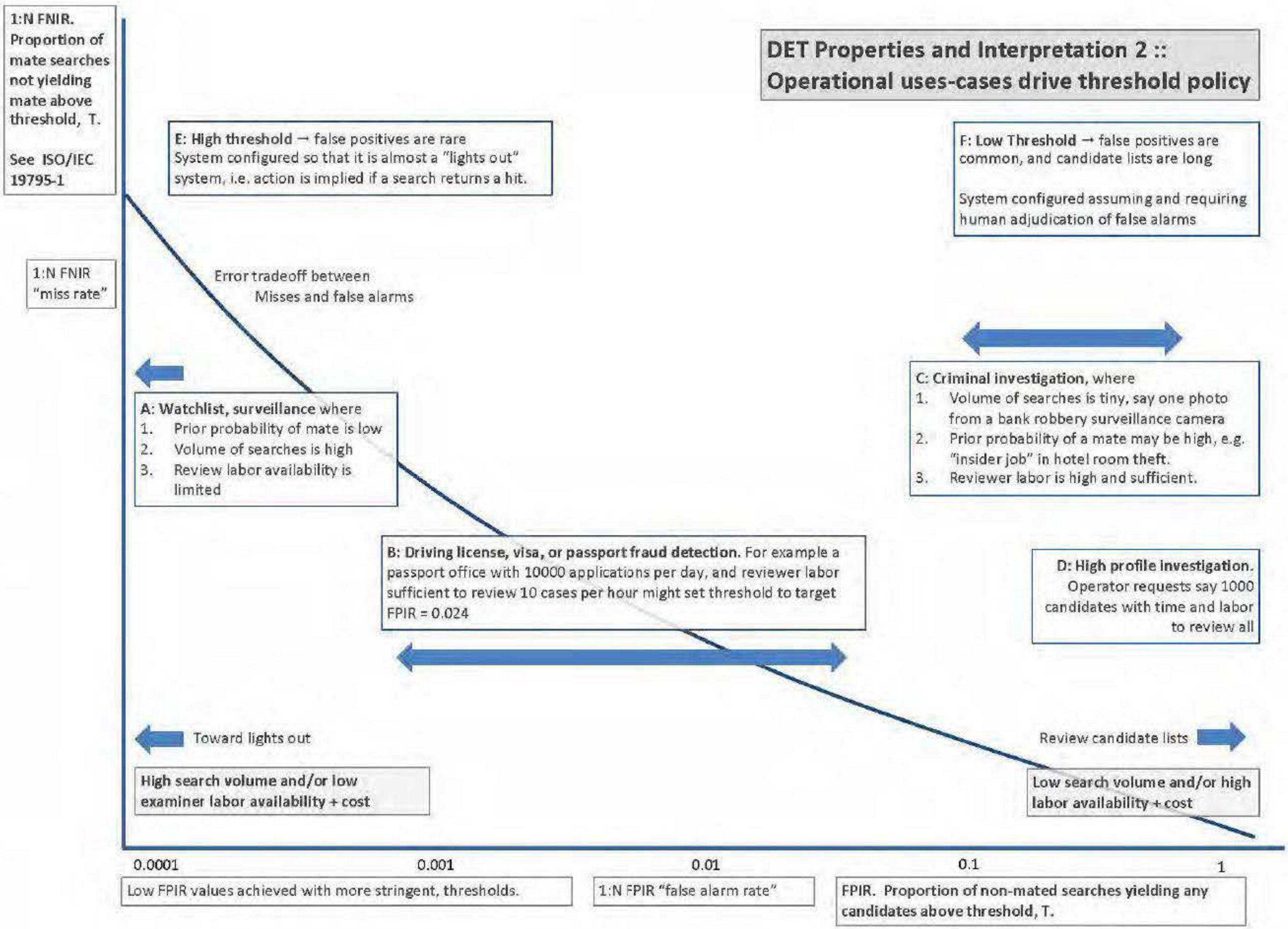


Figure 10: DET as the primary performance reporting mechanism.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

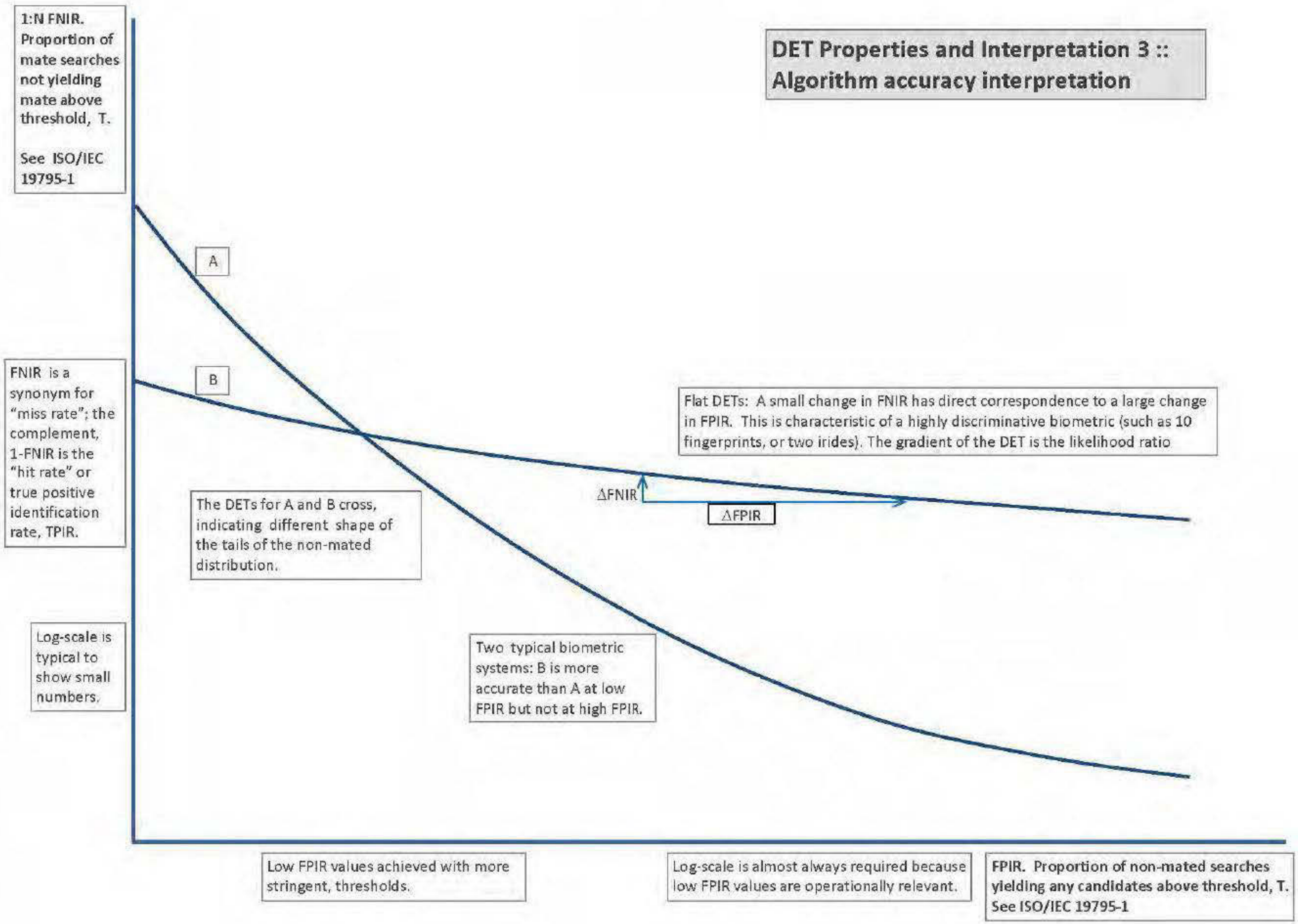


Figure 11: **DET as the primary performance reporting mechanism.**

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

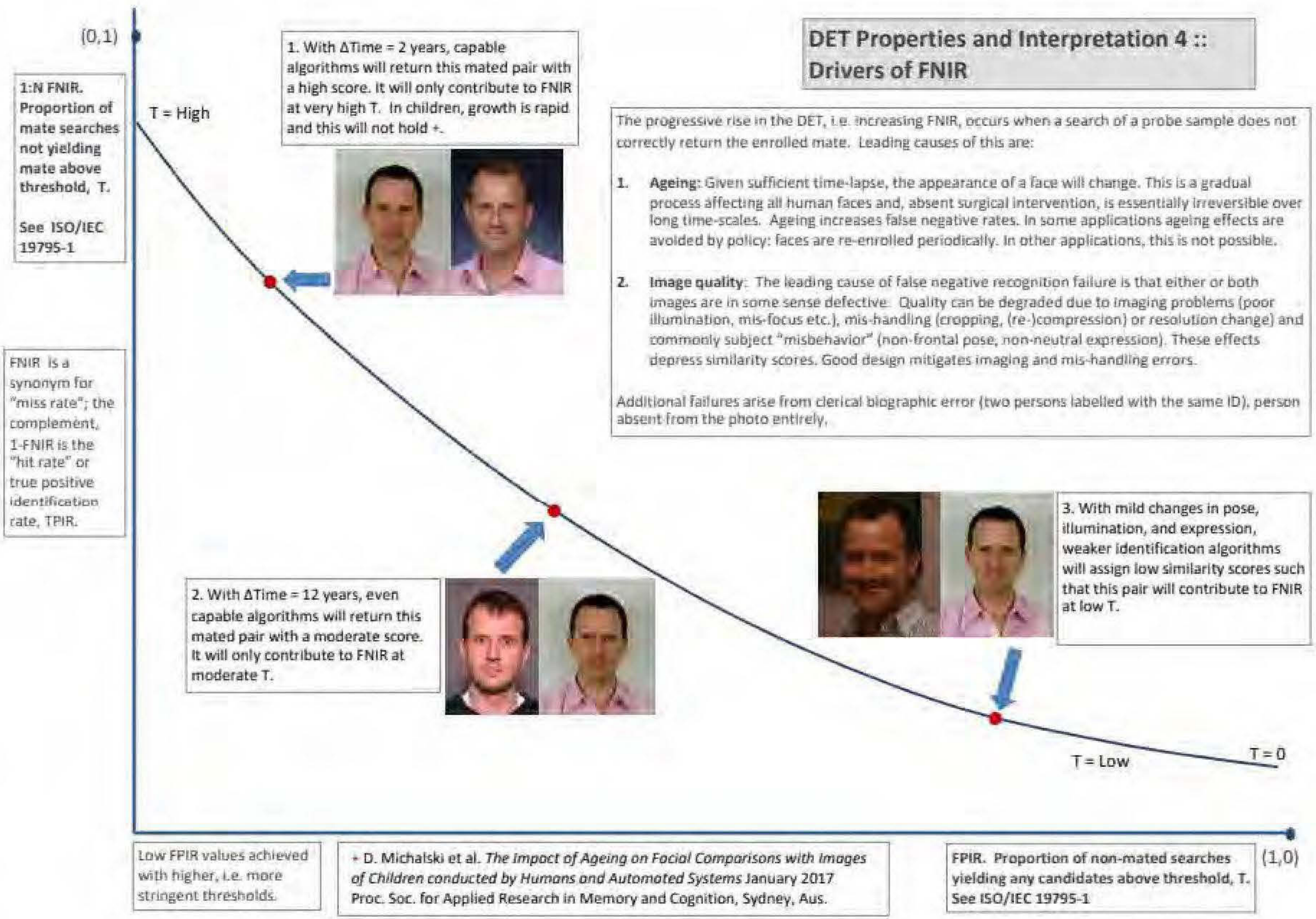


Figure 12: **DET as the primary performance reporting mechanism.**

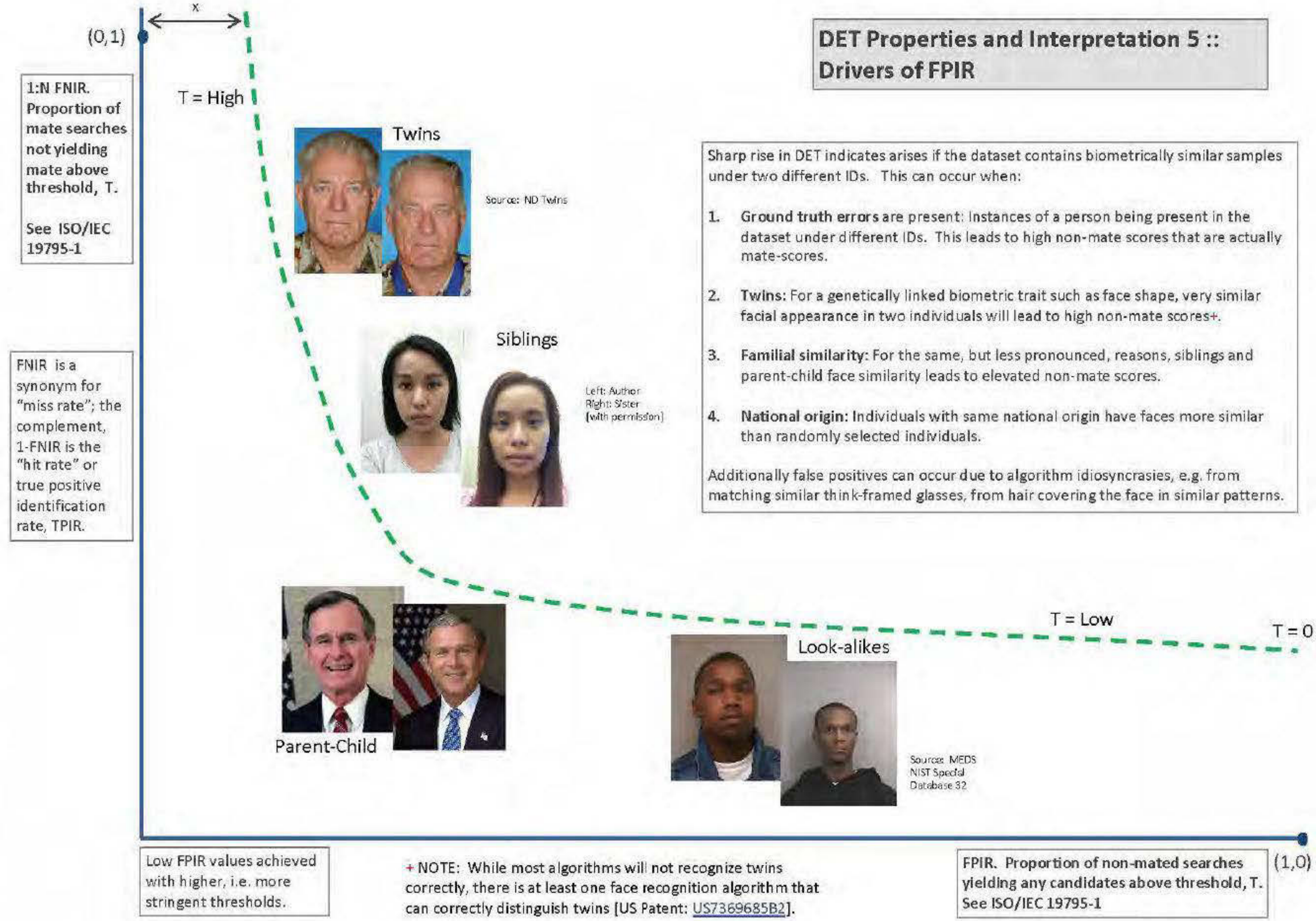


Figure 13: **DET as the primary performance reporting mechanism.**

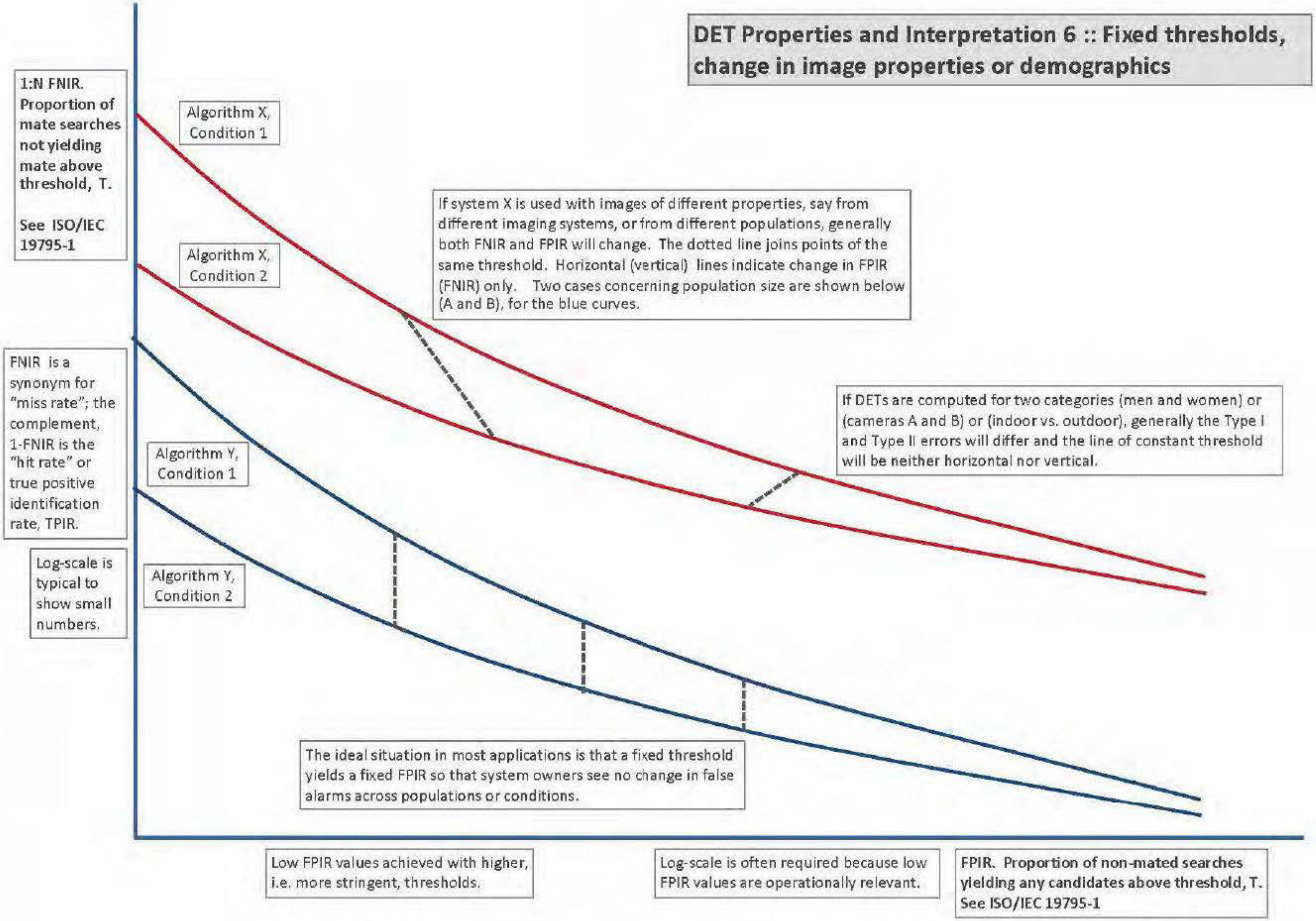


Figure 14: **DET as the primary performance reporting mechanism.**

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

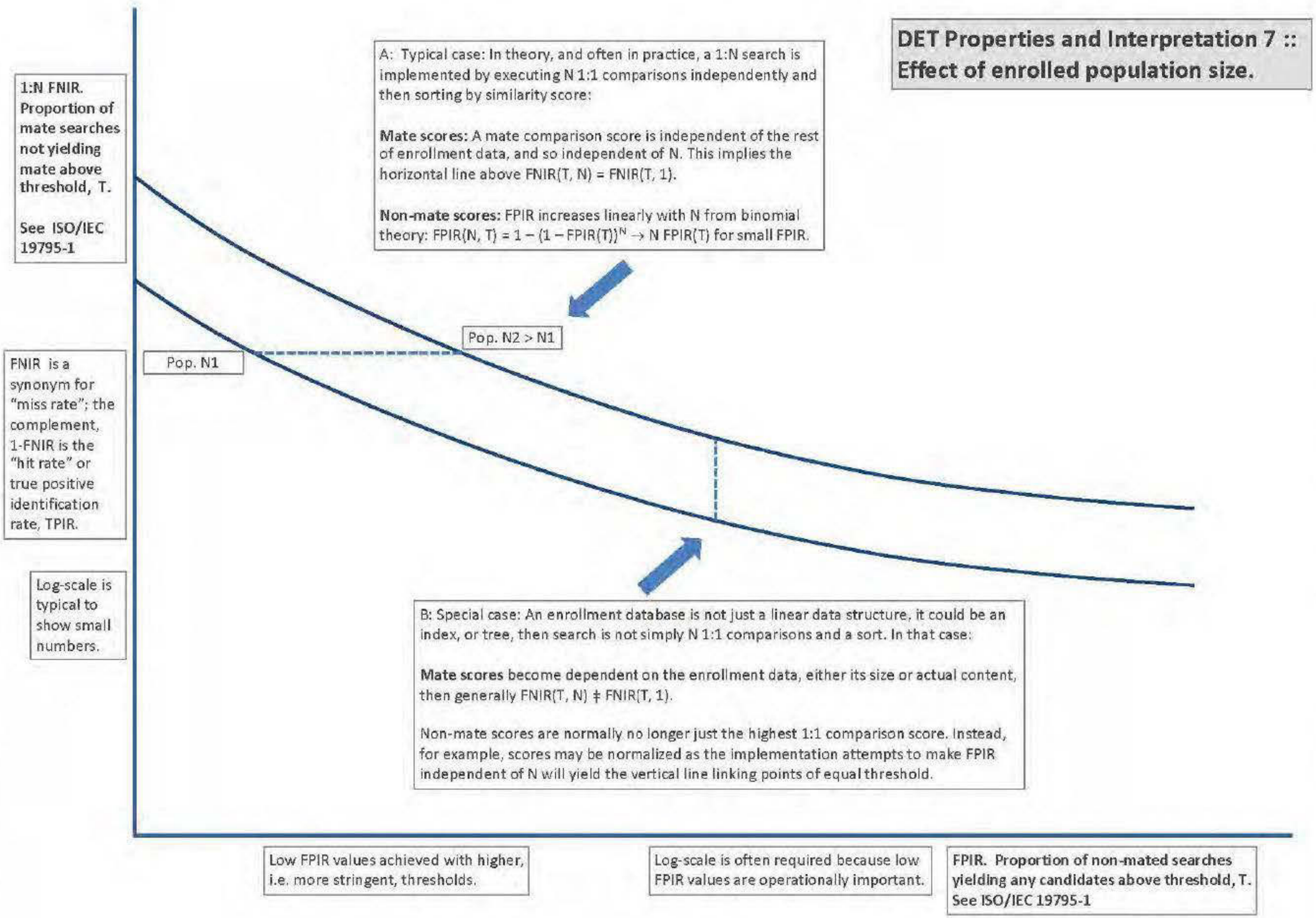


Figure 15: **DET as the primary performance reporting mechanism.**

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

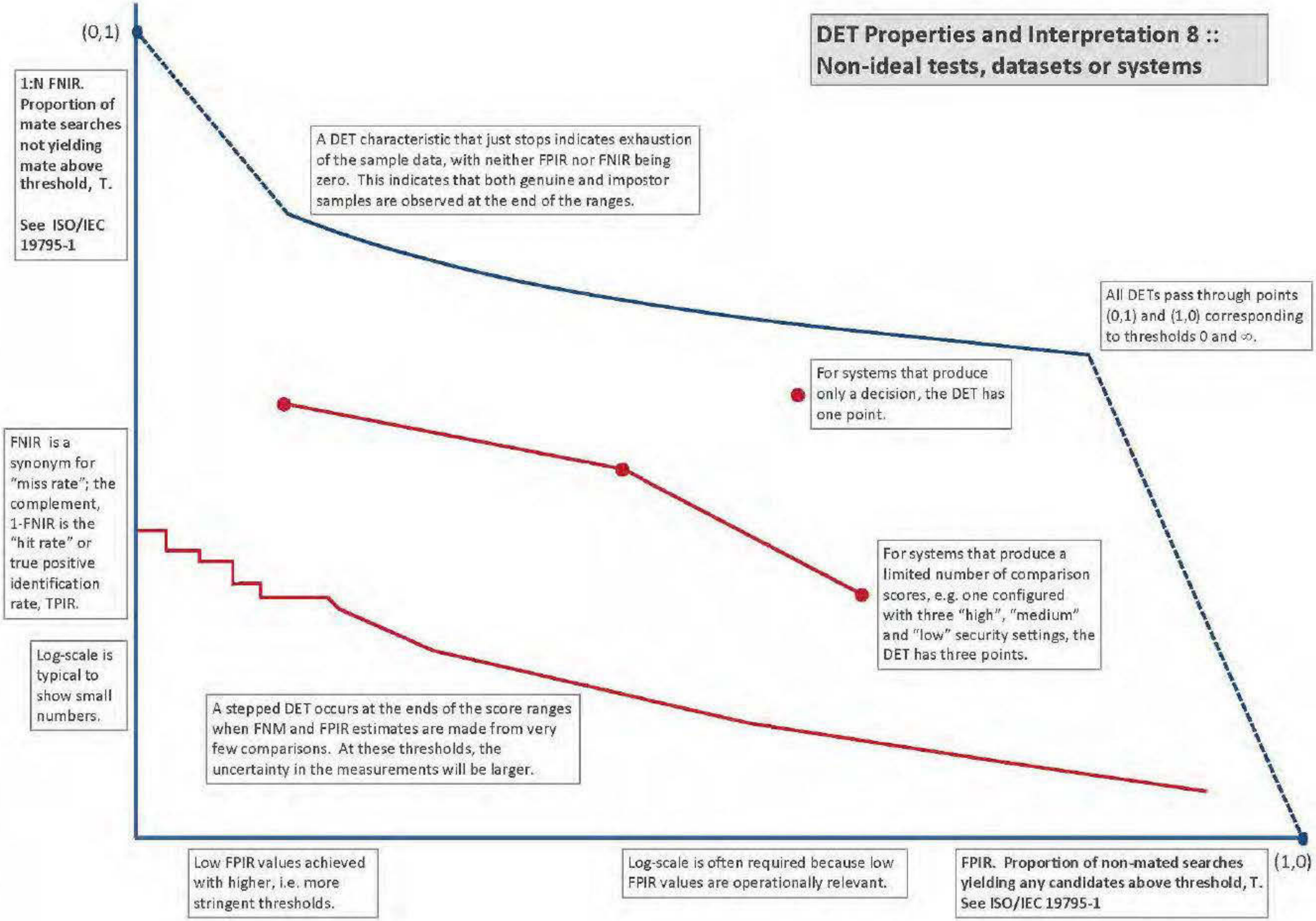


Figure 16: **DET as the primary performance reporting mechanism.**

3.4 Best practice testing requires execution of searches with and without mates

FRVT embeds 1:N searches of two kinds: Those for which there is an enrolled mate, and those for which there is not. The respective numbers for these types of searches appear in Table 5. However, it is common to conduct only mated searches[10]. The cumulative match characteristic is computed from candidate lists produced in mated searches. Even if the CMC is the only metric of interest, the actual trials executed in a test should nevertheless include searches for which no mate exists. As detailed in Table 5 the FRVT reserved disjoint populations of subjects for executing true non-mate searches.

3.5 Failure to extract features

During enrollment some algorithms fail to convert a face image to a template. The proportion of failures is the failure-to-enroll rate, denoted by FTE. Similarly, some search images are not converted to templates. The corresponding proportion is termed failure-to-extract, denoted by FTX.

We do not report FTX because we assume that the same underlying algorithm is used for template generation for enrollment and search.

Failure to extract rates are incorporated into FNIR and FPIR measurements as follows.

- ▷ **Enrollment templates**: Any failed enrollment is regarded as producing a zero length template. Algorithms are required by the API [10] to transparently process zero length templates. The effect of template generation failure on search accuracy depends on whether subsequent searches are mated, or non-mated: Mated searches will fail giving elevated FNIR; non-mated searches will not produce false positives so, to first order, FPIR will be reduced by a factor of $1-\text{FTE}$.
- ▷ **Search templates and 1:N search**: In cases where the algorithm fails to produce a search template from input imagery, the result is taken to be a candidate list whose entries have no hypothesized identities and zero score. The effect of template generation failure on search accuracy depends on whether searches are mated, or non-mated: Mated searches will fail giving elevated FNIR; Non-mated searches will not produce false positives, so FPIR will be reduced. Thus given a measurement of false negative and positive rates made over only those where failures-to-extract did not occur, those rates - call them $\text{FNIR}^\dagger$ and $\text{FPIR}^\dagger$ - could be adjusted by an explicit measurement of FTX as follows

$$\text{FNIR} = \text{FTX} + (1 - \text{FTX})\text{FNIR}^\dagger \tag{8}$$

$$\text{FPIR} = (1 - \text{FTX})\text{FPIR}^\dagger \tag{9}$$

This approach is the correct treatment for positive-identification applications such as access control where cooperative users are enrolled and make attempts at recognition. This approach is not appropriate to negative identification applications, such as visa fraud detection, in which hostile individuals may attempt to evade detection by submitting poor quality samples. In those cases, template generation failures should be investigated as though a false alarm had occurred.

[10] For example, the Megaface benchmark. This is bad practice for several reasons: First, if a developer knows, or can reasonably assume, that a mate always exists, then unrealistic gaming of the test is possible. A second reason is that it does not put FPIR on equal footing with FNIR and that matters because in most applications, not all searches have mates - not everyone has been previously enrolled in a driving license issuance or a criminal justice system - so addressing between-class separation becomes necessary.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

3.6 Fixed length candidate lists, threshold independent workload

Suppose an automated face identification algorithm returns L candidates, and a human reviewer is retained to examine up to R candidates, where $R \leq L$ might be set by policy, preference or labor availability. For now, assume also that the reviewer is not provided with, or ignores, similarity scores, and thresholds are not applied. Given the algorithm typically places mates at low (good) ranks, the number of candidates a reviewer can be expected to review can be derived as follows. Note that the reviewer will:

- ▷ Always inspect the first ranked image — Frac. reviewed = 1
- ▷ Then inspect those candidates where mate not confirmed at rank 1 — Frac. reviewed = 1-CMC(1)
- ▷ Then inspect those candidates where mate not confirmed at rank 1 or 2 — Frac. reviewed = 1-CMC(2)

etc. Thus if the reviewer will stop after a maximum of R candidates, the expected number of candidate reviews is

$$M(R) = 1 + (1 - CMC(1)) + (1 - CMC(2)) + \ldots + (1 - CMC(R-1)) \tag{10}$$

$$= R - \sum_{r=1}^{R-1} CMC(r) \tag{11}$$

A recognition algorithm that front-loads the cumulative match characteristic will offer reduced workload for the reviewer. This workload is defined only over the searches for which a mate exists. In the cases where there truly is no mate, the reviewer would review all R candidates. Thus, if the proportion of searches for which a mate does exist is β, which in the law enforcement context would be the recidivism rate [3], the full expression for workload becomes:

$$M(R) = \beta \left(R - \sum_{r=1}^{R-1} CMC(r) \right) + (1-\beta)R \tag{12}$$

$$= R - \beta \sum_{r=1}^{R-1} CMC(r) \tag{13}$$

3.7 Timing measurement

Algorithms were submitted to NIST as implementations of the application programming interface(API) specified by NIST in the Evaluation Plan [10]. The API includes functions for initialization, template generation, finalization, search, gallery insert, and gallery delete. Two template generation functions are required, one for the preparation of an enrollment template, and one for a search template.

In NIST's test harness, all functions were wrapped by calls to the C++ std::chrono::high resolution clock which on the dedicated timing machine counts 1ns clock ticks. Precision is somewhat worse than that however.

3.8 Uncertainty estimation

3.8.1 Random error

This study leverages operational datasets for measurement of recognition error rates. This affords several advantages. First, large numbers of searches are conducted (see Table 5) giving precision to the measurements. Moreover, for the two mugshot datasets, these do not involve reuse of individuals so binomial statistics can be expected to apply to recognition error counts. In that case, an observed count of a particular recognition outcome (i.e. a false negative or false positive) in M trials will sustain 95% confidence that the actual error rate is no larger than some value.

As an example, the minimum number of mugshot searches conducted in this report is M =154 549, and for an observed FNIR around 0.002, the measurement supports a conclusion that the actual FNIR is no higher than 0.00228 at 99% confidence level. On the false positive side, we tabulate FNIR at FPIR values as low as 0.001. Given estimates based on 331 254 non-mate trials, the actual FPIR values will be below 0.00115 at 99% confidence. In conclusion, large scale evaluation, without reuse of subjects, supports tight uncertainty bounds on the measured error rates.

3.8.2 Systematic error

The FRVT 2018 dataset includes anomalies discovered as a result of inspecting images involved in recognition failures from the most accurate algorithms. Two kinds of failure occur: False negatives (which, for the purpose here, include failures to make templates) and false positives.

False negative errors: We reviewed 600 false negative pairs for which either or both of the leading two algorithms did not put the correct mate in the top 50 candidates. Given 154 549 searches, this number represents 0.39% of the total, resulting in FNIR $\sim$ 0.0039. Of the 600 pairs:

- ▷ **A: Poor quality**: About 20% of the pairs included images of very low quality, often greyscale, low resolution, blurred, low contrast, partially cropped, interlaced, or noisy scans of paper images. Additionally, in a few cases, the face is injured or occluded by bandages or heavy cosmetics.
- ▷ **B: Ground truth identity label bugs**: About 15% of the pairs are not actually mated. We only assigned this outcome when a pair is clearly not mated.
- ▷ **C: Profile views**: About 35% included an image of a profile (side) view of the face, or, more rarely, an image that was rotated 90 degrees in-plane (roll).
- ▷ **D: Tattoos**: About 30% included an image of a tattoo that contained a face image. These arise from mis-labelling in the parent dataset metadata.
- ▷ **E: Ageing**: There is considerable time-lapse between the two captures.

All these estimates are approximate. Of these, the tattoo and mislabled images can never be matched. These constitute an accuracy floor in the sample implying that FNIR cannot be below 0.0018[11]. The profile-views, low-quality images, and images with considerable ageing can, in principle, be successfully matched - indeed some algorithms do so - so are not part of the accuracy floor.

[11] This value is the sum of two partial false negative rates: $\mathrm{FNIR}_B = 0.15 * 0.0039$ plus $\mathrm{FNIR}_D = 0.3 * 0.0039$

For the microsoft-4 algorithm the lowest miss rate from (recent entry in Table 16) is FNIR(640 000, 50, 0) = 0.0018. This is close to the value estimated from the inspection of misses. It is below the 0.0039 figure because the algorithm does match some profile and poor quality images, that the yitu-2 algorithm does not.

For many tables (e.g. Table 16), the FNIR values obtained for the FRVT-2018 mugshots could be corrected by reducing them by 0.0018. The best values would then be indistinct from zero. The results in this report *were not* adjusted to account for this systematic error.

False positive errors: As depicted in Figure 9 many of the DET characteristics in this report exhibit a pronounced turn upward at low false positive rates. The shape can be caused by identity labelling errors in the ground truth of a dataset, specifically persons present in the database under two IDs such that some proportion of non-mate pairs are actually mated. We merged the highest 1000 non-mate pairs produced by three different algorithms which resulted in 1839 unique pairs. This constitutes 0.56% of all non-mate searches. We assert that it is *very* difficult for human reviewers to assign the pairs into the following three categories: twins; doppelgangers; or ground-truth errors (instances of the same person under two IDs). Given this difficulty we made no attempt to correct any ground truth except by removing 57 pairs in the following categories:

- ▷ **A: Profile views**: Thirteen pairs included one or two profile-view images. As described in Figure 102, these can cause false positives.
- ▷ **B: Same-session photographs**: For twelve pairs, the images were identical or trivially altered (e.g. cropped) versions of the same photo. These were present under a different ID likely due to some clerical or procedural mistake.
- ▷ **C: Tattoos of faces**: There were fourteen instances of tattoo photographs that contained faces causing false matches.
- ▷ **D: T-shirt faces**: There were six instances of T-shirt photographs (of Bob Marley and Che Guevara) being detected instead of the face and causing false positives.
- ▷ **E: Background faces**: There were twelve instances of one subject appearing in the background of two otherwise correct portrait photos.

Note we did not remove any images where there was a chance that the pair was actually a different person.

In any case, the results in this report have not been adjusted for this systematic error.

4 Results

This section gives extensive results for algorithms submitted to FRVT 2018. Three page "report cards" for each algorithm are contained in a separate supplement. Performance metrics were described in section 3. The main results are summarized in tabular form with more exhaustive data included as DET, CMC and related graphs in appendices as follows:

- ▷ The three tables 6-8 list algorithms alongside full developer names, acceptance date, size of the provided configuration data, template size and generation time, and search duration data.
  - The **template generation duration** is most important to applications that require fast response. For example, an eGate taking more than two seconds to produce a template might be unacceptable. Note that GPUs may be of utility in expediting this operation for some algorithms, though at additional expense. Two additional factors should be considered[12][13].
  - The **search duration** is the time taken for a search of a search template into a gallery of N enrollment templates. This performance variable, together with the volume of searches, is influential on the amount of hardware needed to sustain an operational deployment. This is measured here with the algorithm running on a single core of a contemporary CPU. Search is most simply implemented as N computations of a distance metric followed by a sort operation to find the closest enrollments. However, considerable optimization of this process is possible, up to and including fast-search algorithms that, by various means, avoid computation of all N distances.
  - The **template size** is the size of the extracted feature vector (or vectors) and any needed header information. Large template sizes may be influential on bus or network bandwidth, storage requirements, and on search duration. While the template itself is an opaque data blob, the feature dimensionality might be estimated by assuming a four-bytes-per-float encoding. There is a wide range of encodings. For the more accurate algorithm, sizes range from 256 bytes to about 2KB bytes, indicating essentially no consensus on face modeling and template design.
  - The **template size multiplier** column shows how, given k input images, the size of the template grows. Most implementations internally extract features from each image and concatenate them, and implement some score-level fusion logic during search. Other implementations, including many of the most accurate algorithms, produce templates whose size does not grow with k. This could be achieved via selection of the best quality image - but this is not optimal in handling ageing where the oldest image could be the best quality. Another mechanism would be feature-level fusion where information is fused from all k inputs. In any case, as a black-box test, the fusion scheme is proprietary and unknown.
  - The size of the **configuration data** is the total size of all files resident in a vendor-provided directory that contains arbitrary read-only files such as parameters, recognition models (e.g caffe). Generally a large value for this quantity may prohibit the use of the algorithm on a resource-constrained device.

[12]The FRVT 2018 API prohibited threading, so some gains from parallelism may be available on multiple-cores or multiple processors, if the feature extraction code could be distributed across them.

[13]Note also that factors of two or more may be realizable by exploiting modern vector processing instructions on CPUs. It is not clear in our measurements whether all developers exploited Intel's AVX2 instructions, for example. Our machine was so equipped, but we insisted that the same compiled library should also run on older machines lacking that instruction. The more sophisticated implementations may have detected AVX2 presence and branched accordingly. The less sophisticated may be defaulted to the reduced instruction set. Readers should see the FRVT 2018 API document for the specific chip details.

- ▷ Tables 16-17 report core rank-based accuracy for mugshot images. The population size is limited to N = 1.6 million identities because this is the largest gallery size on which all algorithms were executed. Notable observations from these tables are as follows:

  - **Accuracy gains during 2018:** NIST Interagency Report 8238 documented massive gains over those reported in the FRVT 2014 report, NIST Interagency Report 8009.

    Further gains are documented in this report. Comparing the most accurate algorithm in June 2018, Microsoft-4, with the most accurate in November 2018, NEC-2, the value of FNIR(N, 1, 0) reduced from 0.0031 to 0.0028 with N = 1.6 million recent images. For lifetime enrollments, Microsoft-4 remained the most accurate algorithm as the newer variants from Microsoft did not reduce this error rate.

    We further note that the revolution is not over: Figure 19 shows that many developers have made great advances in the four months between Phases 1 and 2 of FRVT 2018, Feburary to June. Most developers saw a two-fold reduction in errors, with Neurotechnology seeing a five fold reduction.

  - **Wide range in accuracy:** The rank-1 miss rates vary from FNIR(N, 1, 0) = 0.001 for nec-3 up to about 0.5 for the very fast but inaccurate microfocus-x algorithms. Among the developers who are superior to NEC in 2013, the range is from 0.002 to 0.035 for camvi-3. This large accuracy range is consistent with the buyer-beware maxim, and indicates that face recognition software is far from being commoditized.

- ▷ Tables 19-20 report threshold-based error rates, FNIR(N, L, T), for N = 1.6 million for mugshot-mugshot accuracy on FRVT 2014, FRVT 2018, and also (in pink) mugshot-webcam accuracy using FRVT 2018 enrollments. Notable observations from these tables are as follows:

  - **Order of magnitude accuracy gains since 2014:** As with rank-based results, the gains in accuracy are substantial, though somewhat reduced. At FPIR = 0.01, the best improvement over NEC in 2014 is a nine-fold reduction in FNIR using the Microsoft_4 algorithm. At FPIR = 0.001, the largest gain is a six-fold reduction in FNIR via the Yitu_2 algorithm.

  - **Broad gains across the industry:** About 19 companies realize accuracy better than the NEC benchmark from 2014. This is somewhat lower than the 28 developers who succeeded on the rank-1 metric. This may be due to the ubiquity of, and emphasis on, the rank-1 metric in many published algorithm development papers.

  - **Webcam images:** Searches of webcam images give FNIR(N, T) values around 2 to 3 times higher than mugshot searches. Notably the leading developers with mugshots are approximately the same with poorer quality webcams. But some developers e.g. Camvi, Megvii, TongYi, and Neurotechnology do improve their relative rankings on webcams, perhaps indicating their algorithms were tailored to less constrained images.

- ▷ Tables 10, 12, 13 and show, respectively, high-threshold, rank 1, nd rank 50 FNIR values for all algorithms performing searches into five different gallery sizes, N = 640 000, N = 1 600 000, N = 3 000 000, N = 6 000 000 and 12 000 000. The FPIR = 0.001 table is included to inform high-volume duplicate detection applications. The Rank-1 table is included as a primary accuracy indicator. The Rank-50 table is included to inform agencies who routinely produce 50 candidates for human-review. The notable results are:

  - **Slow growth in rank-based miss rates:** FNIR(N, R) generally grows as a power law, aN^b. From the straight lines of many graphs of Figure 22 this is clearly a reasonable model for most, but not all, algorithms. The coefficient a can be interpreted as FNIR in a gallery of size 1. The more important coefficient b indicates

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

scalability, and often, $b \ll 1$, implies very benign growth in FNIR. The coefficients of the models appear in the Tables 12 and 13.

– **Slow growth in threshold-based miss rates:** FNIR(N, T) also generally grows as a power law, aN^b except at the high threshold values corresponding to low FPIR values. This is visible in the plots of Figure 38 which show straight lines except for FPIR = 0.001, which increase more rapidly with N above 3 000 000. Each trace in those figures shows FNIR(N, T) at fixed FPIR with both N and T varying. Thus at large N, it is usually necessary to elevate T to maintain fixed FPIR. This causes increased FNIR. Why that would no-longer obey a power-law is not known. However, if we expect large galleries to contain individuals with familial relations to the non-mate search images - in the most extreme case, twins - then suppression of false positives becomes more difficult. This is discussed in the Figures starting at Fig. 9

▷ Figure 21 shows false positives from twins against their enrolled siblings, broken out by type of twin: fraternal or identical. The Figure is based on the enrollment of 104 single images on one of a pair of twins, and then the search of 2354 second images. Note that the dataset is heavily skewed towards identical twins which is not representative of the true population. There is also a skew towards same sex fraternal twin pairs compared to different sex fraternal twin pairs again not representative of the true population.

The notable results are:

– For all algorithms tested, the 1087 mated searches (Twin A vs. Twin A) produce scores almost always above typical operational thresholds, with (not shown) matches at rank 1. The images are of good quality, so this is the result expected from the rest of this report.

– For the 1066 identical twin searches (AB), almost all produce the twin at rank 1, with a few producing the mate at further down the candidate lists rank and low score.

– For the 169 fraternal searches (AB) from same sex pairs, most algorithms give a large number of very high scores, implying false positives at all thresholds. However, there there are long tails containing lower scores that are correctly below threshold. In general, scores that are higher in this distribution are all rank 1 whereas the lower scores have much higher ranks.

– (Not shown) Of the 169, there are 24 fraternal searches (AB) involving different sex twins. Here most algorithms correctly report scores well below the lowest threshold, and usually not on the candidate list at all.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

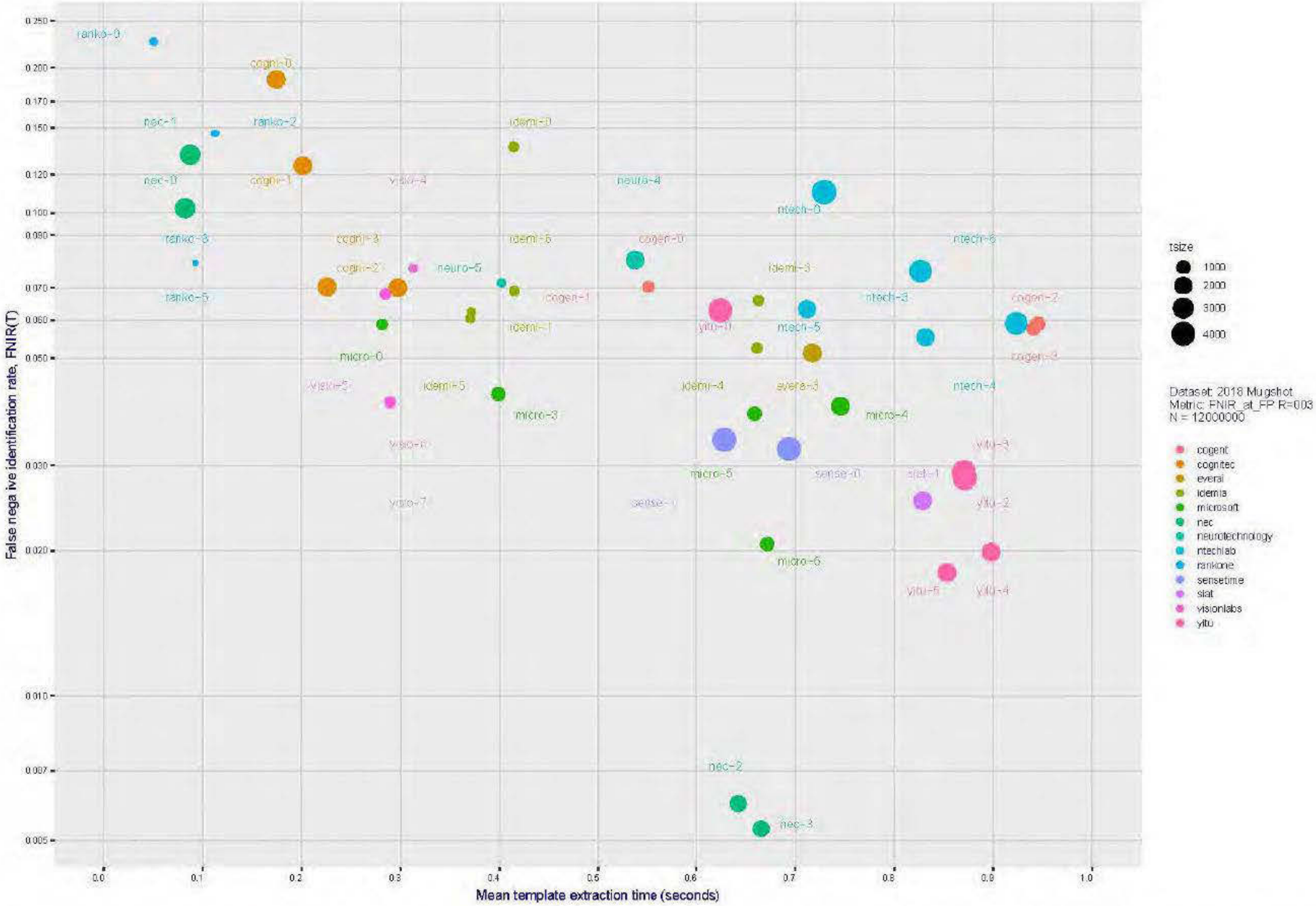


Figure 17: **[Mugshot Dataset] Speed-accuracy tradeoff.** *For developers of the more accurate algorithms the plot shows the tradeoff of high-threshold recognition miss-rates, FNIR(N, N, T) for FPIR(N, T) = 0.003, and template generation time. Developers are coded by color. Template size is encoded by the size of the circle. Some labels are quite distant from the respective point, to avoid superposing text. Without any other influences, the assumption would be that taking time to localize the face, and extract features, would lead to better accuracy. The most notable result, for NEC, is that their slower algorithms are much more accurate than the version that extract features in fewer than 90 milliseconds.*

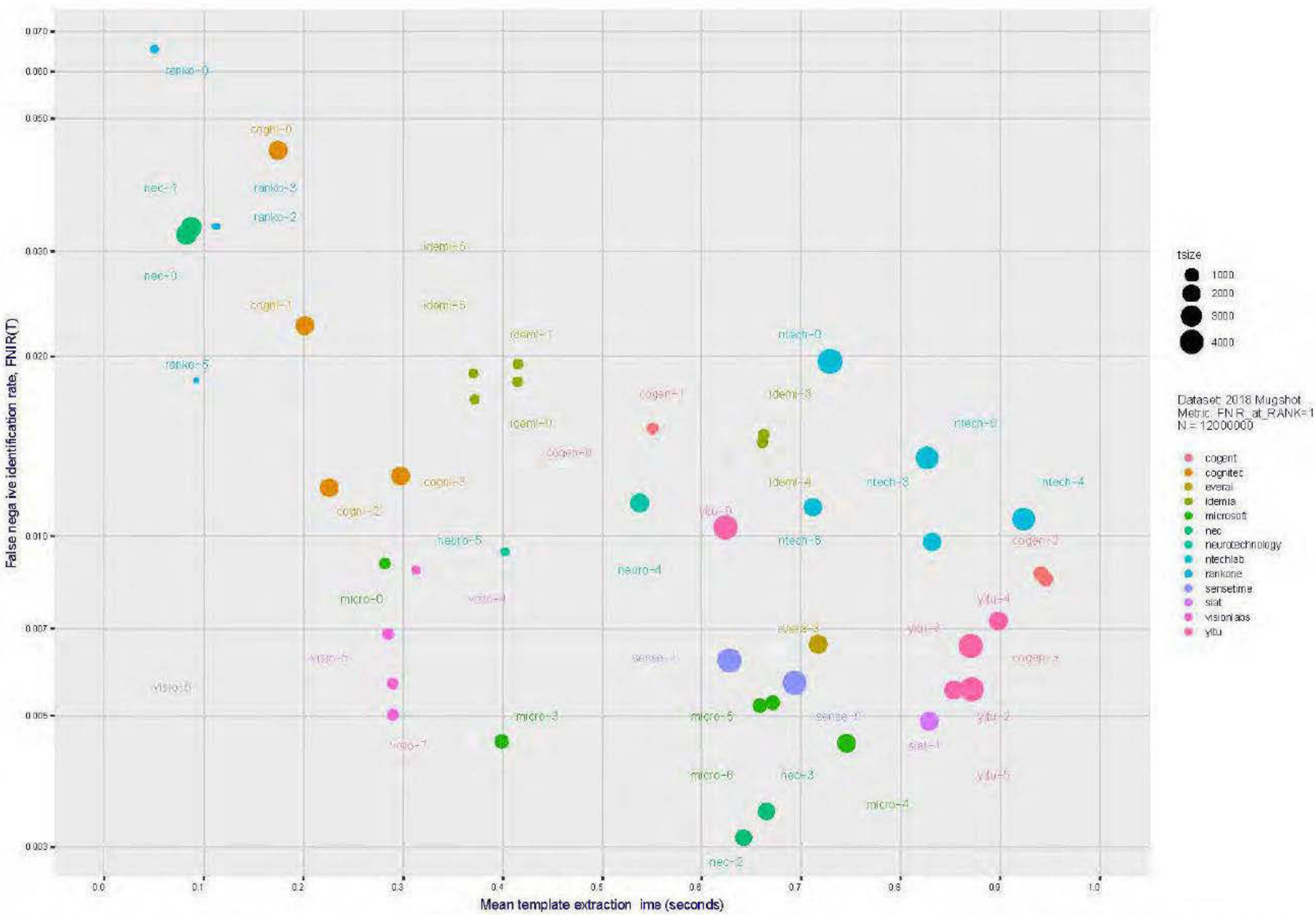


Figure 18: **[Mugshot Dataset] Speed-accuracy tradeoff.** *For developers of the more accurate algorithms the plot shows the tradeoff of rank-one recognition miss-rates, FNIR(N, 1, 0), and template generation time. Developers are coded by color. Template size is encoded by the size of the circle. Some labels are quite distant from the respective point, to avoid superposing text. Without any other influences, the assumption would be that taking time to localize the face, and extract features, would lead to better accuracy. This occurs for NEC with their slower algorithm being much accurate than the version that extract features in fewer than 90 milliseconds.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

	DEVELOPER	SHORT	SEQ.	VALIDATION	CONFIG[1]	TEMPLATE GENERATION			SEARCH DURATION[4] MILLISEC					
	FULL NAME	NAME	NUM.	DATE	DATA (MB)	SIZE(B)	MULT[2]	TIME (MS)[3]	L=1 N=1.6M	L=50 N=1.6M	L=50 N=3M	L=50 N=6M	L=50 N=12M	POWER LAW (μs)
1	3Divi	3divi	0	2018-02-09	186	^{193}4096	k	^{90}426	-	^{187}553	-	-	-	
2	3Divi	3divi	1	2018-02-15	187	^{195}4224	k	^{94}428	-	^{19}37	-	-	-	
3	3Divi	3divi	2	2018-02-15	187	^{47}528	k	^{93}428	-	^{17}33	-	-	-	
4	3Divi	3divi	3	2018-06-19	165	^{41}512	k	^{130}625	^{15}76	^{28}76	-	-	-	
5	3Divi	3divi	4	2018-06-19	186	^{180}4096	k	^{131}628	^{73}604	^{128}801	-	-	-	
6	3Divi	3divi	5	2018-10-26	186	^{178}4096	k	^{138}653	^{67}537	^{104}537	^{51}1376	^{48}2612	^{41}5524	^{71}0.07 $N^{1.1}$
7	3Divi	3divi	6	2018-10-26	187	^{49}528	k	^{141}653	^{10}33	^{15}33	-	-	-	
8	Alchera	alchera	0	2018-06-30	168	^{148}2048	k	^{43}263	^{130}3296	^{193}5420	-	-	-	
9	Alchera	alchera	1	2018-06-30	46	^{124}2048	k	^{8}66	^{121}3516	^{194}5489	-	-	-	
10	Alchera	alchera	2	2018-10-30	7	^{143}2048	k	^{16}115	^{118}2920	^{179}2926	-	-	-	
11	Alchera	alchera	3	2018-10-30	251	^{113}2048	k	^{117}548	^{119}2952	^{191}2953	^{79}6540	^{74}14998	^{70}35227	^{84}0.10 $N^{1.2}$
12	Anke Investments	anke	0	2018-10-30	779	^{165}2072	k	^{96}431	^{74}675	^{123}748	^{53}1482	^{50}2965	^{43}6142	^{58}0.21 $N^{1.1}$
13	Anke Investments	anke	1	2018-10-30	779	^{164}2072	k	^{97}433	^{76}707	^{124}769	-	-	-	
14	Aware	aware	0	2018-02-16	261	^{99}1564	k	^{139}653	-	^{61}251	-	-	-	
15	Aware	aware	1	2018-02-16	232	^{100}1564	k	^{136}651	-	^{61}251	-	-	-	
16	Aware	aware	2	2018-02-16	349	^{167}2076	k	^{197}912	-	^{62}252	-	-	-	
17	Aware	aware	3	2018-06-22	350	^{166}2076	k	^{163}716	^{114}2426	^{174}2508	^{73}4495	-	-	^{41}1.09 $N^{1.0}$
18	Aware	aware	4	2018-06-22	349	^{2}92	k	^{160}712	^{89}1232	^{140}1187	-	-	-	
19	Aware	aware	5	2018-10-30	368	^{173}3100	k	^{182}827	^{18}94	^{36}97	^{13}202	^{11}370	^{9}251	^{11}4.13 $N^{0.7}$
20	Aware	aware	6	2018-10-30	368	^{3}124	k	^{178}818	^{27}167	^{39}162	-	-	-	
21	Ayonix	ayonix	0	2018-06-21	57	^{77}1036	k	^{1}10	^{47}283	^{74}298	-	-	-	
22	Ayonix	ayonix	1	2018-10-29	74	^{80}1036	k	^{3}12	^{44}277	^{70}277	-	-	-	
23	Ayonix	ayonix	2	2018-10-30	74	^{79}1036	1	^{2}11	^{43}277	^{67}274	^{27}531	^{25}1079	^{22}2268	^{80}0.11 $N^{1.0}$
24	Camvi Technologies	camvitech	1	2018-02-16	94	^{69}1024	1	^{24}177	-	^{12}23	-	-	-	
25	Camvi Technologies	camvitech	2	2018-02-16	442	^{74}1024	1	^{172}774	-	^{11}20	-	-	-	
26	Camvi Technologies	camvitech	3	2018-06-30	233	^{72}1024	1	^{158}707	^{7}10	^{9}11	-	-	-	
27	Camvi Technologies	camvitech	4	2018-10-30	233	^{66}1024	1	^{165}718	^{11}33	^{14}32	^{8}38	^{6}40	^{4}48	^{3}8492.66 $N^{0.1}$
28	Camvi Technologies	camvitech	5	2018-10-30	257	^{61}1024	1	^{170}769	^{9}31	^{13}30	-	-	-	
29	Thales	cogent	0	2018-06-20	533	^{46}525	k	^{118}551	^{63}494	^{110}558	^{42}1047	^{41}2060	^{38}4141	^{21}0.46 $N^{1.0}$
30	Thales	cogent	1	2018-06-20	533	^{45}525	k	^{119}552	^{64}498	^{108}556	^{43}1048	^{42}2082	^{35}4263	^{30}0.39 $N^{1.0}$
31	Thales	cogent	2	2018-10-30	681	^{84}1043	k	^{208}987	^{106}2017	^{166}2144	^{73}4298	^{69}8472	^{65}16429	^{37}1.08 $N^{1.0}$
32	Thales	cogent	3	2018-10-30	681	^{83}1043	k	^{202}960	^{88}1230	^{146}1311	^{63}2687	^{60}5398	^{58}10184	^{39}0.62 $N^{1.0}$
33	Cognitec Systems GmbH	cognitec	0	2018-06-21	364	^{155}2052	k	^{23}176	^{100}1748	^{154}1780	^{68}3672	^{64}7093	^{63}15224	^{55}0.57 $N^{1.0}$
34	Cognitec Systems GmbH	cognitec	1	2018-06-21	412	^{149}2052	k	^{28}202	^{103}1835	^{156}1805	^{71}3971	^{67}7484	^{64}16249	^{60}0.49 $N^{1.1}$
35	Cognitec Systems GmbH	cognitec	2	2018-10-30	463	^{151}2052	k	^{34}227	^{99}1733	^{153}1763	^{67}3660	^{66}7279	^{59}13895	^{40}0.83 $N^{1.0}$
36	Cognitec Systems GmbH	cognitec	3	2018-10-30	465	^{157}2052	k	^{52}297	^{98}1719	^{155}1791	^{66}3638	^{65}7277	^{61}14904	^{52}0.66 $N^{1.0}$
37	Dahua Technology Co. Ltd	dahua	0	2018-10-29	276	^{131}2048	k	^{73}378	-	^{65}256	-	-	-	
38	Dahua Technology Co. Ltd	dahua	1	2018-10-29	276	^{118}2048	k	^{68}371	-	^{64}256	^{33}601	^{31}1199	^{30}3001	^{78}0.02 $N^{1.2}$
39	Dermalog	dermalog	0	2018-02-16	0	^{5}128	1	^{65}344	-	^{88}404	-	-	-	
40	Dermalog	dermalog	1	2018-02-16	0	^{8}128	1	^{22}171	-	^{91}407	-	-	-	
41	Dermalog	dermalog	2	2018-02-16	0	^{18}256	k	^{64}344	-	^{119}640	-	-	-	
42	Dermalog	dermalog	3	2018-06-21	0	^{7}128	1	^{31}211	^{17}92	^{34}92	-	-	-	
43	Dermalog	dermalog	4	2018-06-21	0	^{4}128	1	^{29}208	^{16}91	^{35}93	-	-	-	
44	Dermalog	dermalog	5	2018-10-26	0	^{6}128	1	^{109}532	^{2}0	^{1}0	^{1}0	^{1}0	^{1}0	^{4}66.21 $N^{0.3}$
45	Dermalog	dermalog	6	2018-10-26	0	^{24}256	1	^{108}514	^{25}141	^{35}143	^{18}267	^{16}527	^{14}1285	^{83}0.05 $N^{1.0}$
46	Ever AI	everai	0	2018-06-21	142	^{140}2048	1	^{99}438	^{4}4	^{5}3	^{2}5	-	-	^{9}42.41 $N^{0.3}$
47	Ever AI	everai	1	2018-06-21	200	^{111}2048	1	^{125}590	^{51}336	^{81}356	^{35}651	-	-	^{74}0.03 $N^{1.1}$
48	Ever AI	everai	2	2018-10-30	224	^{132}2048	1	^{71}377	^{46}278	^{72}283	-	-	-	
49	Ever AI	everai	3	2018-10-30	438	^{112}2048	1	^{167}735	^{45}278	^{71}281	^{30}572	^{29}1146	^{23}2278	^{48}0.12 $N^{1.0}$
50	Eyedea Recognition	eyedea	0	2018-02-16	644	^{194}4152	k	^{97}424	-	^{120}640	-	-	-	
51	Eyedea Recognition	eyedea	1	2018-02-16	287	^{82}1036	k	^{56}311	-	^{78}307	-	-	-	
52	Eyedea Recognition	eyedea	2	2018-02-16	287	^{78}1036	k	^{85}429	-	^{75}305	-	-	-	

Notes	
1	Configuration size does not capture static data present in libraries. Libraries are not counted because most implementations include common ancillary libraries for image processing (e.g. openCV) or numerical computation (e.g. blas).
2	This multiplier expresses the increase in template size when k images are passed to the template generation function.
3	All durations are measured on Intel®Xeon®CPU E5-2630 v4 @ 2.20GHz processors. Estimates are made by wrapping the API function call in calls to std::chrono::high_resolution_clock which on the machine in (3) counts 1ns clock ticks. Precision is somewhat worse than that however.
4	Search durations are measured as in the prior note. The power-law model in the final column mostly fits the empirical results in Figure 111. However in certain cases the model is not correct and should not be used numerically.

Table 6: Summary of algorithms and properties included in this report. The blue superscripts give ranking for the quantity in that column. Missing search durations, denoted by "-", are absent because those runs were not executed, usually because we did not run on the larger galleries. Caution: The power-law model is sometimes an incorrect model. It is included here only to show broad sublinear behavior, which is flagged in green. The models should not be used for prediction.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

	DEVELOPER FULL NAME	SHORT NAME	SEQ. NUM.	VALIDATION DATE	CONFIG[1] DATA (MB)	TEMPLATE GENERATION SIZE (B)	MULT[2]	TIME (MS)[3]	SEARCH DURATION[4] MILLISEC L=1 N=1.6M	L=50 N=1.6M	L=50 N=3M	L=50 N=6M	L=50 N=12M	POWER LAW (µs)
53	Eyedea Recognition	eyedea	3	2018-06-18	284	^{80}1036	k	^{75}385	^{48}309	^{78}311	-	-	-	
54	Glory Ltd	glory	0	2018-06-30	0	^{33}418	k	^{18}160	^{70}575	^{112}575	-	-	-	
55	Glory Ltd	glory	1	2018-06-30	0	^{108}1726	k	^{81}405	^{104}1864	^{158}1978	-	-	-	
56	Gorilla Technology	gorilla	0	2018-02-01	95	^{202}8300	k	^{91}427	-	^{200}10426	-	-	-	
57	Gorilla Technology	gorilla	1	2018-06-19	91	^{170}2156	k	^{21}169	^{128}5254	^{190}5156	-	-	-	
58	Gorilla Technology	gorilla	2	2018-10-29	91	^{87}1132	k	^{62}341	^{26}145	^{27}146	^{19}293	^{17}612	^{17}1509	^{66}0.02 $N^{1.1}$
59	Gorilla Technology	gorilla	3	2018-10-26	94	^{169}2156	k	^{124}563	^{105}1924	^{161}2047	-	-	-	
60	loginface Corp	hbinno	0	2018-02-01	88	^{44}520	-	^{43}265	-	^{95}419	-	-	-	
61	Hikvision Research Institute	hikvision	0	2018-02-12	378	^{106}1808	1	^{174}875	-	^{171}2360	-	-	-	
62	Hikvision Research Institute	hikvision	1	2018-02-12	378	^{107}1808	1	^{178}820	-	^{172}2403	-	-	-	
63	Hikvision Research Institute	hikvision	2	2018-02-12	378	^{106}1808	1	^{176}820	-	^{173}2408	-	-	-	
64	Hikvision Research Institute	hikvision	3	2018-06-30	408	^{91}1408	1	^{133}633	^{84}904	^{138}1108	^{60}2377	^{53}3785	^{45}7570	^{20}0.91 $N^{1.0}$
65	Hikvision Research Institute	hikvision	4	2018-06-30	334	^{86}1152	1	^{104}510	^{79}784	^{134}1024	^{58}2094	^{52}3254	^{44}7117	^{19}0.86 $N^{1.0}$
66	Hikvision Research Institute	hikvision	5	2018-10-29	593	^{90}1408	-	^{129}619	^{83}883	^{132}895	^{55}1908	^{54}3792	^{52}9387	^{72}0.10 $N^{1.1}$
67	Hikvision Research Institute	hikvision	6	2018-10-29	593	^{89}1408	1	^{126}610	^{82}871	^{131}877	-	-	-	
68	Idemia	idemia	0	2018-02-16	371	^{32}364	1	^{86}416	-	^{28}133	^{14}249	^{12}502	-	^{33}0.08 $N^{1.0}$
69	Idemia	idemia	1	2018-02-16	371	^{30}364	1	^{87}417	-	^{33}138	-	-	-	
70	Idemia	idemia	2	2018-02-16	371	^{31}364	1	^{88}417	-	^{34}138	-	-	-	
71	Idemia	idemia	3	2018-06-21	472	^{48}528	1	^{149}689	^{50}318	^{82}361	^{34}631	^{28}1104	^{24}2332	^{12}5.03 $N^{0.8}$
72	Idemia	idemia	4	2018-06-21	472	^{50}528	1	^{147}689	^{29}168	^{53}211	^{25}475	^{23}995	^{21}2225	^{73}0.02 $N^{1.1}$
73	Idemia	idemia	5	2018-10-29	417	^{28}352	1	^{70}374	^{20}137	^{32}138	^{23}437	^{19}724	^{19}1630	^{82}0.01 $N^{1.2}$
74	Idemia	idemia	6	2018-10-29	417	^{29}352	1	^{69}373	^{21}137	^{31}138	^{24}442	^{22}827	^{20}1646	^{83}0.01 $N^{1.2}$
75	Imagus Technology Pty Ltd	imagus	0	2018-02-14	35	^{37}512	k	^{5}43	-	^{48}202	-	-	-	
76	Imagus Technology Pty Ltd	imagus	2	2018-06-21	35	^{34}512	k	^{9}76	^{37}200	^{52}208	-	-	-	
77	Imagus Technology Pty Ltd	imagus	3	2018-06-21	46	^{39}512	k	^{7}57	^{38}201	^{50}206	-	-	-	
78	Incode Technologies	incode	0	2018-06-29	23	^{76}1024	k	^{27}190	^{93}1293	^{186}3510	-	-	-	
79	Incode Technologies	incode	1	2018-06-29	151	^{144}2048	k	^{151}690	^{94}1542	^{188}4497	-	-	-	
80	Incode Technologies	incode	2	2018-10-29	71	^{120}2048	1	^{45}291	^{59}411	^{89}404	-	-	-	
81	Incode Technologies	incode	3	2018-10-29	133	^{128}2048	1	^{156}704	^{58}408	^{94}412	^{38}846	^{35}1606	^{36}4482	^{69}0.05 $N^{1.1}$
82	Innovatrics	innovatrics	0	2018-02-16	0	^{53}530	k	^{110}455	-	^{118}625	-	-	-	
83	Innovatrics	innovatrics	1	2018-02-16	0	^{51}530	k	^{58}316	-	^{117}625	-	-	-	
84	Innovatrics	innovatrics	2	2018-06-21	0	^{52}530	k	^{40}255	^{3}1	^{3}2	-	-	-	
85	Innovatrics	innovatrics	3	2018-06-21	0	^{54}530	k	^{41}255	^{109}2020	^{157}1882	-	-	-	
86	Innovatrics	innovatrics	4	2018-10-30	0	^{55}1076	k	^{83}406	^{6}8	^{8}8	^{4}11	^{3}9	^{3}13	^{7}668.38 $N^{0.8}$
87	Alivia / Innovation Sys.	isystems	0	2018-02-14	262	^{134}2048	1	^{33}222	-	^{65}393	-	-	-	
88	Alivia / Innovation Sys.	isystems	1	2018-02-14	263	^{68}1024	1	^{32}222	-	^{55}240	-	-	-	
89	Alivia / Innovation Sys.	isystems	2	2018-06-25	268	^{126}2048	1	^{59}316	^{55}385	^{99}484	^{50}1275	^{37}1770	^{31}3063	^{16}0.68 $N^{0.9}$
90	Alivia / Innovation Sys.	isystems	3	2018-10-30	350	^{142}2048	1	^{159}856	^{54}384	^{64}387	^{41}976	^{40}1817	^{51}9319	^{98}0.00 $N^{1.3}$
91	Lookman Electroplast Industries	lookman	3	2018-10-28	203	^{26}292	1	^{63}342	^{77}739	^{122}745	^{52}1394	^{48}2817	^{46}8286	^{64}0.13 $N^{1.1}$
92	Lookman Electroplast Industries	lookman	4	2018-10-28	184	^{56}548	1	^{60}325	^{89}981	^{133}998	-	-	-	
93	Megvii	megvii	0	2018-02-15	1327	^{138}2048	1	^{174}794	-	^{73}284	^{26}530	^{24}1060	-	^{30}0.18 $N^{1.0}$
94	Megvii	megvii	1	2018-10-28	1703	^{184}4096	1	^{137}652	^{68}551	^{111}560	^{49}1219	^{45}2316	^{42}5956	^{68}0.08 $N^{1.1}$
95	Megvii	megvii	2	2018-10-28	1735	^{182}4096	1	^{142}656	^{69}552	^{109}557	-	-	-	
96	Microfocus	microfocus	0	2018-02-12	101	^{21}256	k	^{106}525	-	^{42}184	-	-	-	
97	MicroFocus	microfocus	1	2018-02-16	101	^{12}256	k	^{107}527	-	^{20}39	-	-	-	
98	Microfocus	microfocus	2	2018-02-16	101	^{22}256	k	^{108}529	-	^{4}2	-	-	-	
99	Microfocus	microfocus	3	2018-06-22	101	^{14}256	k	^{46}269	^{33}185	^{45}188	-	-	-	
100	Microfocus	microfocus	4	2018-06-22	102	^{20}256	k	^{47}270	^{34}186	^{46}189	-	-	-	
101	Microfocus	microfocus	5	2018-10-29	94	^{25}256	k	^{45}266	^{31}182	^{44}186	^{20}353	^{18}706	^{18}1422	^{34}0.11 $N^{1.0}$
102	Microfocus	microfocus	6	2018-10-29	94	^{18}256	k	^{44}265	^{32}182	^{43}186	-	-	-	
103	Microsoft	microsoft	0	2018-01-30	126	^{43}512	1	^{48}283	-	^{114}593	^{47}1193	^{46}2295	^{38}4936	^{31}0.22 $N^{1.0}$
104	Microsoft	microsoft	1	2018-02-12	165	^{88}1024	1	^{66}349	-	^{130}869	-	-	-	

Notes	
1	Configuration size does not capture static data present in libraries. Libraries are not counted because most implementations include common ancillary libraries for image processing (e.g. openCV) or numerical computation (e.g. blas).
2	This multiplier expresses the increase in template size when k images are passed to the template generation function.
3	All durations are measured on Intel®Xeon®CPU E5-2630 v4 @ 2.20GHz processors. Estimates are made by wrapping the API function call in calls to std::chrono::high_resolution_clock which on the machine in (3) counts 1ns clock ticks. Precision is somewhat worse than that however.
4	Search durations are measured as in the prior note. The power-law model in the final column mostly fits the empirical results in Figure 111. However in certain cases the model is not correct and should not be used numerically.

Table 7: Summary of algorithms and properties included in this report. The blue superscripts give ranking for the quantity in that column. Missing search durations, denoted by "-", are absent because those runs were not executed, usually because we did not run on the larger galleries. Caution: The power-law model is sometimes an incorrect model. It is included here only to show broad sublinear behavior, which is flagged in green. The models should not be used for prediction.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

	DEVELOPER FULL NAME	SHORT NAME	SEQ. NUM.	VALIDATION DATE	CONFIG[1] DATA (MB)	TEMPLATE GENERATION SIZE (B)	MULT[2]	TIME (MS)[3]	SEARCH DURATION[4] MILLISEC L=1 N=1.6M	L=50 N=1.6M	L=50 N=3M	L=50 N=6M	L=50 N=12M	POWER LAW (μs)
105	Microsoft	microsoft	2	2018-02-12	228	^{73}1024	1	^{130}555	-	^{129}869	-	-	-	
106	Microsoft	microsoft	3	2018-06-20	230	^{65}1024	1	^{80}404	^{96}1638	^{148}1603	^{65}3260	^{60}6730	^{58}13833	^{86}0.51 $N^{1.1}$
107	Microsoft	microsoft	4	2018-06-20	487	^{127}2048	1	^{171}773	^{117}2662	^{177}2691	^{75}5260	^{71}11070	^{67}22748	^{87}0.83 $N^{1.1}$
108	Microsoft	microsoft	5	2018-10-29	381	^{70}1024	1	^{148}673	^{95}1604	^{150}1671	^{64}3073	^{61}6296	^{57}13147	^{88}0.79 $N^{1.0}$
109	Microsoft	microsoft	6	2018-10-29	478	^{62}1024	1	^{153}695	^{97}1640	^{149}1617	^{69}3707	^{62}6394	^{56}12879	^{47}0.68 $N^{1.0}$
110	NEC	nec	0	2018-06-21	131	^{172}2592	k	^{10}82	^{49}317	^{98}426	^{37}738	^{33}1315	^{27}2737	^{14}0.73 $N^{0.9}$
111	NEC	nec	1	2018-06-29	131	^{171}2592	k	^{11}88	^{36}193	^{51}208	^{22}388	^{20}750	^{18}1577	^{18}0.21 $N^{1.0}$
112	NEC	nec	2	2018-10-30	705	^{101}1616	k	^{140}653	^{57}405	^{93}409	^{44}1072	^{37}1755	^{34}4255	^{70}0.06 $N^{1.1}$
113	NEC	nec	3	2018-10-30	774	^{102}1712	k	^{150}690	^{5}7	^{7}7	^{5}14	^{5}40	^{6}82	^{80}0.00 $N^{1.2}$
114	Neurotechnology	neurotech	0	2018-02-16	331	^{197}5214	k	^{154}702	-	^{182}3040	-	-	-	
115	Neurotechnology	neurotech	1	2018-02-16	331	^{198}5214	k	^{145}661	-	^{184}3054	-	-	-	
116	Neurotechnology	neurotech	2	2018-02-16	331	^{199}5214	k	^{144}658	-	^{183}3051	-	-	-	
117	Neurotechnology	neurotech	3	2018-06-27	265	^{116}2048	k	^{116}547	^{87}1094	^{135}1059	^{59}2111	^{57}4779	^{49}8798	^{31}0.73 $N^{1.0}$
118	Neurotechnology	neurotech	4	2018-06-27	265	^{145}2048	k	^{115}543	^{86}1060	^{136}1061	^{57}2091	^{56}4263	^{47}8736	^{17}1.22 $N^{1.0}$
119	Neurotechnology	neurotech	5	2018-10-30	266	^{17}256	k	^{84}412	^{80}835	^{127}839	^{54}1690	^{51}3219	^{50}8955	^{62}0.19 $N^{1.1}$
120	Neurotechnology	neurotech	6	2018-10-30	564	^{16}256	k	^{169}746	^{81}839	^{128}842	-	-	-	
121	Newland Computer Co. Ltd	newland	2	2018-10-30	96	^{110}2048	-	^{191}868	^{134}8653	^{199}8765	^{96}17713	^{81}38963	-	^{67}1.32 $N^{1.1}$
122	Noblis	noblis	1	2018-10-30	114	^{128}2048	1	^{30}211	^{91}1273	^{143}1272	-	-	-	
123	Noblis	noblis	2	2018-10-30	153	^{200}6144	1	^{110}535	^{116}2513	^{175}2522	^{76}5649	^{72}12432	^{73}44262	^{85}0.04 $N^{1.3}$
124	N-Tech Lab	ntech	0	2018-02-16	2124	^{196}4442	k	^{166}730	-	^{83}382	^{36}673	^{34}1344	-	^{22}0.27 $N^{1.0}$
125	N-Tech Lab	ntech	1	2018-02-16	851	^{104}1736	k	^{82}405	-	^{38}161	-	-	-	
126	N-Tech Lab	ntech	3	2018-06-21	3664	^{174}3484	k	^{184}831	^{53}384	^{80}326	^{31}596	^{30}1192	^{25}2411	^{34}0.24 $N^{1.0}$
127	N-Tech Lab	ntech	4	2018-06-21	3766	^{175}3484	k	^{198}929	^{52}379	^{79}312	^{32}597	^{32}1204	^{26}2416	^{28}0.21 $N^{1.0}$
128	N-Tech Lab	ntech	5	2018-10-30	1685	^{108}1940	k	^{164}717	^{42}243	^{57}246	^{28}538	^{26}1100	^{28}2867	^{78}0.02 $N^{1.1}$
129	N-Tech Lab	ntech	6	2018-10-30	1686	^{109}1940	k	^{187}841	^{41}243	^{56}246	^{29}546	^{27}1104	^{29}2873	^{77}0.02 $N^{1.1}$
130	Quantasoft	quantasoft	1	2018-10-30	276	^{117}2048	k	^{76}396	^{135}15422	^{201}14858	^{95}14717	-	^{66}18323	
131	Rank One Computing	rankone	0	2018-02-07	0	^{12}228	k	^{6}50	-	^{22}75	^{11}142	^{10}220	^{10}502	^{15}0.12 $N^{0.9}$
132	Rank One Computing	rankone	1	2018-02-15	0	^{27}324	k	^{17}136	-	^{41}169	-	-	-	
133	Rank One Computing	rankone	2	2018-06-19	0	^{10}133	k	^{14}113	^{22}138	^{29}137	^{16}258	^{14}517	^{12}1029	^{35}0.10 $N^{1.0}$
134	Rank One Computing	rankone	3	2018-06-19	0	^{11}133	k	^{15}114	^{23}138	^{30}137	^{15}258	^{13}515	^{11}1027	^{36}0.09 $N^{1.0}$
135	Rank One Computing	rankone	4	2018-10-09	0	^{1}85	k	^{4}36	^{19}101	^{27}101	^{12}190	-	-	^{27}0.07 $N^{1.0}$
136	Rank One Computing	rankone	5	2018-10-24	0	^{9}133	k	^{12}94	^{24}140	^{36}144	^{17}266	^{15}525	^{13}1049	^{33}0.11 $N^{1.0}$
137	RealNetworks	realnetworks	0	2018-06-21	96	^{185}4100	1	^{38}244	^{123}4257	^{178}2740	-	-	-	
138	RealNetworks	realnetworks	1	2018-06-21	105	^{189}4104	k	^{37}243	^{122}3568	^{164}2107	-	-	-	
139	RealNetworks	realnetworks	2	2018-10-30	105	^{187}4104	k	^{39}245	^{107}2008	^{160}2046	^{72}4190	^{70}8633	^{61}15020	^{36}1.08 $N^{1.0}$
140	KanKan AI	remarkai	0	2018-10-30	187	^{129}2048	k	^{127}615	^{131}5685	^{195}5723	-	-	-	
141	KanKan AI	remarkai	1	2018-10-30	187	^{114}2048	k	^{98}434	^{130}5680	^{196}5761	^{84}12475	^{80}28726	^{76}59618	^{81}0.27 $N^{1.2}$
142	Sensetime Group Ltd	sensetime	0	2018-10-30	525	^{186}4104	k	^{162}715	^{65}498	^{100}501	^{48}1212	^{43}2281	^{40}5032	^{65}0.09 $N^{1.1}$
143	Sensetime Group Ltd	sensetime	1	2018-10-30	525	^{188}4104	k	^{143}656	^{66}516	^{101}502	^{45}1146	^{44}2301	^{37}4765	^{68}0.09 $N^{1.1}$
144	Shaman Software	shaman	0	2018-02-12	0	^{181}4096	k	^{113}538	-	^{102}523	-	-	-	
145	Shaman Software	shaman	1	2018-02-12	0	^{179}4096	k	^{121}557	-	^{103}524	-	-	-	
146	Shaman Software	shaman	2	2018-02-12	0	^{201}8192	k	^{122}557	-	^{121}688	-	-	-	
147	Shaman Software	shaman	3	2018-06-30	0	^{125}2048	k	^{155}704	^{75}692	^{77}310	-	-	-	
148	Shaman Software	shaman	4	2018-06-30	0	^{126}2048	k	^{135}642	^{61}434	^{66}267	-	-	-	
149	Shaman Software	shaman	6	2018-10-26	0	^{133}2048	k	^{157}706	^{72}594	^{115}608	-	-	-	
150	Shaman Software	shaman	7	2018-10-26	0	^{123}2048	k	^{158}709	^{71}593	^{116}605	^{46}1169	^{47}2411	^{35}5007	^{49}0.25 $N^{1.0}$
151	Shenzhen Inst. Adv. Tech. CAS	SIAT	0	2018-02-14	306	^{86}1096	k	^{67}358	-	^{147}1343	-	-	-	
152	Shenzhen Inst. Adv. Tech. CAS	SIAT	1	2018-06-30	521	^{147}2052	1	^{188}842	^{125}4512	^{186}4402	^{81}9103	^{76}18391	^{71}28745	^{44}2.06 $N^{1.0}$
153	Shenzhen Inst. Adv. Tech. CAS	SIAT	2	2018-02-30	521	^{153}2052	1	^{195}906	^{126}6101	^{189}4884	^{87}9556	^{77}18834	^{73}39717	^{45}2.08 $N^{1.0}$
154	Smilart	smilart	0	2018-02-15	105	^{64}1024	k	^{31}168	-	^{144}1285	-	-	-	
155	Smilart	smilart	1	2018-02-15	120	^{71}1024	k	^{146}662	-	^{139}1135	-	-	-	
156	Smilart	smilart	2	2018-02-15	109	^{67}1024	k	^{123}560	-	^{145}1302	-	-	-	

Notes	
1	Configuration size does not capture static data present in libraries. Libraries are not counted because most implementations include common ancillary libraries for image processing (e.g. openCV) or numerical computation (e.g. blas).
2	This multiplier expresses the increase in template size when k images are passed to the template generation function.
3	All durations are measured on Intel®Xeon®CPU E5-2630 v4 @ 2.20GHz processors. Estimates are made by wrapping the API function call in calls to std::chrono::high_resolution_clock which on the machine in (3) counts 1ns clock ticks. Precision is somewhat worse than that however.
4	Search durations are measured as in the prior note. The power-law model in the final column mostly fits the empirical results in Figure 111. However in certain cases the model is not correct and should not be used numerically.

Table 8: Summary of algorithms and properties included in this report. The blue superscripts give ranking for the quantity in that column. Missing search durations, denoted by "-", are absent because those runs were not executed, usually because we did not run on the larger galleries. Caution: The power-law model is sometimes an incorrect model. It is included here only to show broad sublinear behavior, which is flagged in green. The models should not be used for prediction.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

	DEVELOPER FULL NAME	SHORT NAME	SEQ. NUM.	VALIDATION DATE	CONFIG[1] DATA (MB)	TEMPLATE GENERATION SIZE (B)	MULT[2]	TIME (MS)[3]	SEARCH DURATION[4] MILLISEC L=1 N=1.6M	L=50 N=1.6M	L=50 N=3M	L=50 N=6M	L=50 N=12M	POWER LAW (μs)
157	Smilart	smilart	4	2018-10-30	65	^{36}512	k	^{19}167	^{136}15879	^{210}15382	-	-	-	
158	Smilart	smilart	5	2018-10-30	562	^{110}2048	k	^{101}464	-	-	-	-	-	
159	Synesis	synesis	0	2018-02-15	332	^{36}512	k	^{36}237	-	^{40}162	-	-	-	
160	Synesis	synesis	3	2018-10-30	237	^{177}4096	k	^{13}103	^{78}784	^{125}796	^{56}1928	^{56}3861	^{48}8748	^{76}0.07 $N^{1.1}$
161	Tevian	tevian	0	2018-02-16	666	^{112}2048	1	^{75}394	-	^{80}405	-	-	-	
162	Tevian	tevian	1	2018-02-16	666	^{119}2048	1	^{79}398	-	^{87}403	-	-	-	
163	Tevian	tevian	2	2018-02-16	666	^{115}2048	1	^{77}397	-	^{86}402	-	-	-	
164	Tevian	tevian	3	2018-06-20	707	^{118}2048	1	^{54}300	^{62}473	^{106}539	-	-	-	
165	Tevian	tevian	4	2018-06-20	707	^{111}2048	1	^{53}299	^{60}434	^{105}537	-	-	-	
166	Tevian	tevian	5	2018-10-30	773	^{113}2048	1	^{83}416	^{56}405	^{92}407	^{37}852	^{36}1753	^{36}3373	^{54}0.14 $N^{1.0}$
167	TigerIT Americas LLC	tiger	0	2018-06-29	333	^{132}2052	k	^{95}428	^{102}1822	^{180}2942	-	-	-	
168	TigerIT Americas LLC	tiger	1	2018-06-27	333	^{149}2052	k	^{78}398	^{1}0	^{1}1	-	-	-	
169	TigerIT Americas LLC	tiger	2	2018-10-29	416	^{136}2052	k	^{103}464	^{101}1814	^{158}1919	^{70}3829	^{68}7519	^{60}14805	^{43}0.83 $N^{1.0}$
170	TigerIT Americas LLC	tiger	3	2018-10-30	416	^{134}2052	k	^{102}464	^{35}191	^{47}189	-	-	-	
171	TongYi Transportation Technology	tongyi	0	2018-06-29	1701	^{160}2070	k	^{26}190	^{118}2256	^{169}2272	-	-	-	
172	TongYi Transportation Technology	tongyi	1	2018-06-29	1701	^{161}2070	1	^{25}189	^{117}2238	^{168}2257	-	-	-	
173	Toshiba	toshiba	0	2018-10-30	961	^{98}1548	k	^{200}930	^{133}6147	^{197}6230	^{83}12209	^{79}25330	^{75}49398	^{79}0.36 $N^{1.2}$
174	Toshiba	toshiba	1	2018-10-30	961	^{159}2060	k	^{201}931	^{132}6001	^{196}6549	-	-	-	
175	Visidon	visidon	0	2018-06-20	208	^{76}1028	k	^{61}337	^{106}2006	^{176}2566	-	-	-	
176	Visidon	visidon	1	2018-10-30	166	^{130}2052	k	^{159}695	^{124}4357	^{187}4458	^{80}8429	^{75}17210	^{69}34185	^{55}2.40 $N^{1.0}$
177	Vigilant Solutions	vigilant	0	2018-02-08	335	^{95}1544	k	^{180}823	-	^{162}2058	-	-	-	
178	Vigilant Solutions	vigilant	1	2018-02-14	249	^{136}2056	k	^{166}739	-	^{163}2075	-	-	-	
179	Vigilant Solutions	vigilant	2	2018-02-14	335	^{97}1544	k	^{177}820	-	^{166}2121	-	-	-	
180	Vigilant Solutions	vigilant	3	2018-06-21	335	^{96}1544	k	^{185}832	^{115}2453	^{170}2307	-	-	-	
181	Vigilant Solutions	vigilant	4	2018-06-21	337	^{90}1544	k	^{183}830	^{110}2050	^{167}2251	-	-	-	
182	Vigilant Solutions	vigilant	5	2018-10-30	335	^{96}1544	k	^{173}778	-	^{152}1720	-	-	-	
183	Vigilant Solutions	vigilant	6	2018-10-30	337	^{91}1544	k	^{186}834	-	^{181}1713	-	-	-	
184	VisionLabs	visionlabs	3	2018-02-16	624	^{15}256	1	^{35}228	-	^{6}5	^{3}5	^{2}6	-	^{6}417.87 $N^{0.2}$
185	VisionLabs	visionlabs	4	2018-06-22	299	^{22}256	1	^{57}315	^{8}19	^{10}17	^{6}20	^{4}26	^{3}29	^{3}2663.29 $N^{0.1}$
186	VisionLabs	visionlabs	5	2018-06-22	305	^{31}512	1	^{55}300	^{13}54	^{16}33	^{7}37	^{8}56	^{7}88	^{10}166.84 $N^{0.4}$
187	VisionLabs	visionlabs	6	2018-10-30	360	^{40}512	1	^{50}292	^{12}36	^{18}36	^{9}39	^{7}44	^{5}53	^{9}3211.93 $N^{0.2}$
188	VisionLabs	visionlabs	7	2018-10-30	360	^{41}512	1	^{51}293	^{14}63	^{21}63	^{10}72	^{9}80	^{8}115	^{8}2076.32 $N^{0.2}$
189	Vocord	vocord	0	2018-02-16	872	^{56}608	k	^{111}536	-	^{67}268	-	-	-	
190	Vocord	vocord	1	2018-02-16	872	^{57}608	k	^{112}536	-	^{68}268	-	-	-	
191	Vocord	vocord	2	2018-02-16	924	^{119}2048	k	^{134}635	-	^{59}248	-	-	-	
192	Vocord	vocord	3	2018-06-30	627	^{89}896	k	^{161}714	^{29}215	^{58}247	-	-	-	
193	Vocord	vocord	4	2018-06-30	627	^{90}896	k	^{114}538	^{30}216	^{63}253	-	-	-	
194	Vocord	vocord	5	2018-10-30	1035	^{58}768	k	^{179}822	^{28}158	^{49}204	^{21}383	^{21}767	^{16}1466	^{32}0.12 $N^{1.0}$
195	Vocord	vocord	6	2018-10-30	1035	^{203}10240	k	^{181}825	^{30}170	^{54}216	-	-	-	
196	Zhuhai Yisheng Electronics Tech.	yisheng	0	2018-02-14	473	^{166}2108	k	^{129}615	-	^{113}587	-	-	-	
197	Zhuhai Yisheng Electronics Tech.	yisheng	1	2018-06-19	474	^{176}3704	k	^{74}387	^{111}2228	^{137}1108	-	-	-	
198	Shanghai Yitu Technology	yitu	0	2018-02-12	1774	^{191}4136	1	^{132}633	-	^{98}464	^{40}868	^{38}1769	-	^{59}0.12 $N^{1.1}$
199	Shanghai Yitu Technology	yitu	1	2018-02-12	1944	^{190}4136	1	^{199}920	-	^{97}463	-	-	-	
200	Shanghai Yitu Technology	yitu	2	2018-06-21	2077	^{193}4138	1	^{192}870	^{129}5516	^{192}5417	^{77}6101	^{73}13264	^{68}33047	^{14}9.25 $N^{0.9}$
201	Shanghai Yitu Technology	yitu	3	2018-06-21	2077	^{192}4138	1	^{193}871	^{127}5248	^{191}5242	^{78}6286	^{78}19829	^{74}45621	^{61}1.08 $N^{1.1}$
202	Shanghai Yitu Technology	yitu	4	2018-10-30	2119	^{162}2070	1	^{198}910	^{92}1288	^{142}1203	^{61}2440	^{59}5241	^{54}9671	^{46}0.52 $N^{1.0}$
203	Shanghai Yitu Technology	yitu	5	2018-10-30	2043	^{160}2070	1	^{190}861	^{90}1235	^{141}1197	^{62}2508	^{58}5003	^{53}9601	^{42}0.55 $N^{1.0}$

Notes	
1	Configuration size does not capture static data present in libraries. Libraries are not counted because most implementations include common ancillary libraries for image processing (e.g. openCV) or numerical computation (e.g. blas).
2	This multiplier expresses the increase in template size when k images are passed to the template generation function.
3	All durations are measured on Intel(R)Xeon(R)CPU E5-2630 v4 @ 2.20GHz processors. Estimates are made by wrapping the API function call in calls to std::chrono::high_resolution_clock which on the machine in (3) counts 1ns clock ticks. Precision is somewhat worse than that however.
4	Search durations are measured as in the prior note. The power-law model in the final column mostly fits the empirical results in Figure 111. However in certain cases the model is not correct and should not be used numerically.

Table 9: Summary of algorithms and properties included in this report. The blue superscripts give ranking for the quantity in that column. Missing search durations, denoted by "-", are absent because those runs were not executed, usually because we did not run on the larger galleries. Caution: The power-law model is sometimes an incorrect model. It is included here only to show broad sublinear behavior, which is flagged in green. The models should not be used for prediction.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

MISSES BELOW THRESHOLD, T		ENROL LIFETIME					ENROL MOST RECENT				
FNIR(N, T > 0, R > L)		DATASET: FRVT 2018					DATASET: FRVT 2018				
#	ALGORITHM	N=0.64M	N=1.6M	N=3.0M	N=6.0M	N=12.0M	N=0.64M	N=1.6M	N=3.0M	N=6.0M	N=12.0M
1	3DIVI-3	^{139}0.3000	^{124}0.3499	^{62}0.3859	^{59}0.4344		^{146}0.3550	^{140}0.4023			
2	3DIVI-5	^{95}0.1045	^{94}0.1339				^{101}0.1382	^{107}0.1691	^{77}0.1938	^{68}0.2392	^{67}0.3087
3	ALCHERA-0	^{85}0.0852	^{86}0.1105	^{50}0.1361	^{48}0.1913		^{96}0.1128	^{96}0.1405			
4	ALCHERA-3	^{92}0.1018	^{92}0.1296				^{98}0.1205	^{98}0.1590	^{70}0.1891	^{69}0.2467	^{72}0.3628
5	ANKE-0	^{78}0.0769	^{77}0.0989				^{84}0.0968	^{93}0.1199	^{66}0.1432	^{68}0.1811	^{60}0.2624
6	AWARE-3	^{84}0.0846	^{75}0.0991	^{47}0.1148	^{43}0.1459		^{94}0.1122	^{79}0.1306	^{67}0.1471	^{62}0.1793	^{57}0.2395
7	AWARE-5	^{131}0.2628	^{113}0.2984				^{144}0.3459	^{139}0.3729	^{80}0.4094	^{77}0.4615	^{88}0.2637
8	AYONIX-0	^{171}0.8262	^{139}0.8490	^{67}0.8640	^{62}0.8809		^{182}0.7795	^{180}0.8114			
9	AYONIX-2	^{168}0.7602	^{137}0.8038				^{181}0.7867	^{182}0.8246	^{86}0.8511	^{81}0.8708	^{79}0.8946
10	CAMVI-3	^{35}0.0281	^{46}0.0509	^{39}0.0680	^{47}0.1871		^{41}0.0415	^{56}0.0736			
11	CAMVI-4	^{29}0.0257	^{47}0.0505				^{38}0.0395	^{57}0.0741	^{51}0.1008	^{70}0.2532	^{64}0.2731
12	COGENT-0	^{51}0.0387	^{45}0.0434	^{29}0.0523	^{26}0.0784	^{13}0.1569	^{52}0.0455	^{45}0.0557	^{40}0.0734	^{43}0.1194	^{40}0.2029
13	COGENT-1	^{68}0.0598	^{49}0.0518				^{51}0.0455	^{44}0.0557	^{41}0.0734	^{42}0.1194	^{39}0.2029
14	COGENT-2	^{19}0.0220	^{18}0.0299	^{35}0.0390	^{23}0.0703	^{16}0.1595	^{14}0.0356	^{30}0.0475	^{31}0.0655	^{41}0.1185	^{46}0.2241
15	COGENT-3	^{30}0.0258	^{27}0.0341	^{24}0.0450	^{29}0.0842	^{25}0.1864	^{27}0.0361	^{38}0.0515	^{42}0.0771	^{50}0.1374	^{56}0.2488
16	COGNITEC-0	^{91}0.0989	^{90}0.1256				^{102}0.1400	^{99}0.1628	^{71}0.1592	^{68}0.2205	^{66}0.2859
17	COGNITEC-1	^{67}0.0597	^{66}0.0777	^{41}0.0946	^{40}0.1315	^{38}0.2552	^{77}0.0832	^{77}0.1045	^{53}0.1244	^{53}0.1561	^{51}0.2338
18	COGNITEC-2	^{41}0.0296	^{36}0.0401	^{28}0.0523	^{31}0.0852	^{34}0.2298	^{48}0.0433	^{46}0.0560	^{35}0.0695	^{33}0.0980	^{36}0.1967
19	COGNITEC-3	^{38}0.0288	^{38}0.0397	^{27}0.0505	^{28}0.0837	^{32}0.2140	^{49}0.0427	^{43}0.0555	^{32}0.0679	^{31}0.0938	^{29}0.1840
20	DAHUA-1	^{54}0.0410	^{50}0.0521				^{60}0.0596	^{59}0.0755	^{47}0.0905	^{40}0.1179	^{33}0.1910
21	DERMALOG-4	^{142}0.3405	^{128}0.3892	^{64}0.4181	^{60}0.4533		^{154}0.4380	^{153}0.4813			
22	DERMALOG-5	^{62}0.0490	^{62}0.0649				^{74}0.0726	^{71}0.0909	^{56}0.1172	^{58}0.1618	^{59}0.2516
23	DERMALOG-6	^{36}0.0276	^{37}0.0283				^{42}0.0420	^{45}0.0542	^{34}0.0687	^{37}0.1004	^{28}0.1812
24	EVERAI-0	^{57}0.0460	^{66}0.0676				^{66}0.0681	^{73}0.0921	^{57}0.1223		
25	EVERAI-1	^{28}0.0255	^{39}0.0360				^{33}0.0383	^{47}0.0518	^{33}0.0686		
26	EVERAI-3	^{15}0.0191	^{15}0.0256	^{11}0.0338	^{8}0.0389		^{17}0.0282	^{17}0.0377	^{16}0.0473	^{19}0.0683	^{27}0.1653
27	EYEDEA-3	^{137}0.2911	^{122}0.3283	^{61}0.3673	^{58}0.4154		^{148}0.3496	^{142}0.3893			
28	GLORY-1	^{129}0.2160	^{110}0.2447	^{56}0.2618	^{53}0.2884		^{136}0.2790	^{133}0.3067			
29	GORILLA-2	^{100}0.1088	^{96}0.1379				^{106}0.1561	^{108}0.1902	^{76}0.2210	^{71}0.2625	^{71}0.3426
30	HIK-2	^{101}0.1104	^{99}0.1363	^{51}0.1610	^{49}0.2061	^{41}0.3067	^{89}0.0985	^{88}0.1212			
31	HIK-3	^{86}0.0885	^{85}0.1097				^{78}0.0853	^{78}0.1054	^{58}0.1228	^{54}0.1552	^{57}0.2500
32	HIK-4	^{83}0.0839	^{83}0.1031	^{48}0.1225	^{46}0.1518	^{39}0.2618	^{76}0.0821	^{74}0.1013	^{56}0.1173	^{53}0.1498	^{58}0.2503
33	HIK-5	^{18}0.0218	^{22}0.0308	^{18}0.0397	^{22}0.0661		^{32}0.0339	^{27}0.0467	^{26}0.0583	^{32}0.0967	^{44}0.2164
34	IDEMIA-0	^{70}0.0645	^{65}0.0802	^{42}0.0986	^{39}0.1237	^{26}0.1872	^{81}0.0920	^{81}0.1135	^{62}0.1332	^{59}0.1628	^{45}0.2208
35	IDEMIA-1	^{49}0.0304	^{38}0.0377	^{28}0.0465	^{18}0.0623	^{14}0.1578	^{47}0.0444	^{40}0.0540	^{29}0.0647	^{26}0.0856	^{22}0.1618
36	IDEMIA-2	^{58}0.0453	^{58}0.0564	^{32}0.0668	^{33}0.0896	^{20}0.1706	^{49}0.0449	^{42}0.0543			
37	IDEMIA-3	^{23}0.0238	^{21}0.0308				^{31}0.0373	^{35}0.0497	^{48}0.0927	^{73}0.2387	^{74}0.4442
38	IDEMIA-4	^{20}0.0223	^{14}0.0276	^{10}0.0338	^{11}0.0478	^{11}0.1556	^{19}0.0326	^{19}0.0399	^{17}0.0472	^{27}0.0644	^{26}0.1659
39	IDEMIA-5	^{33}0.0261	^{24}0.0319	^{17}0.0395	^{15}0.0588	^{22}0.1764	^{34}0.0385	^{26}0.0465	^{30}0.0562	^{25}0.0788	^{31}0.1951
40	IDEMIA-6	^{26}0.0253	^{25}0.0316	^{14}0.0383	^{14}0.0581	^{29}0.2046	^{32}0.0377	^{24}0.0458	^{30}0.0550	^{22}0.0760	^{47}0.2242
41	IMAGUS-2	^{164}0.6616	^{130}0.7143	^{58}0.7503	^{61}0.7867		^{177}0.7092	^{176}0.7510			
42	INCODE-1	^{107}0.1400	^{108}0.1796	^{54}0.2159	^{52}0.2741		^{114}0.1763	^{114}0.2143			
43	INCODE-3	^{89}0.0949	^{88}0.1227				^{110}0.1349	^{108}0.1703	^{73}0.1986	^{67}0.2379	^{58}0.3157
44	INNOVATRICS-4	^{82}0.0837	^{74}0.0928				^{95}0.1106	^{84}0.1340	^{68}0.1418	^{52}0.1418	^{11}0.1418
45	ISYSTEMS-0	^{61}0.0485	^{61}0.0633	^{36}0.0795	^{37}0.1057	^{30}0.2072	^{70}0.0707	^{72}0.0912			
46	ISYSTEMS-1	^{59}0.0480	^{60}0.0627	^{35}0.0784	^{36}0.1054	^{30}0.2081	^{69}0.0702	^{69}0.0903			
47	ISYSTEMS-2	^{52}0.0394	^{53}0.0545	^{34}0.0679			^{62}0.0612	^{62}0.0814	^{50}0.1006	^{51}0.1405	^{52}0.2374
48	ISYSTEMS-3	^{42}0.0301	^{41}0.0402	^{31}0.0557	^{32}0.0881	^{28}0.1992	^{58}0.0464	^{52}0.0620	^{45}0.0940	^{46}0.1324	^{54}0.2417
49	LOOKMAN-3	^{46}0.0335	^{43}0.0425				^{30}0.0372	^{25}0.0463	^{20}0.0541	^{21}0.0758	^{24}0.1650
50	MEGVII-0	^{81}0.0822	^{82}0.1023	^{49}0.1228	^{44}0.1489	^{35}0.2348	^{80}0.0895	^{80}0.1086	^{61}0.1287	^{57}0.1606	^{49}0.2288
51	MEGVII-1						^{55}0.0586	^{58}0.0746	^{46}0.0896	^{47}0.1338	^{65}0.2761
52	MICROFOCUS-3	^{179}0.9002	^{141}0.9213	^{69}0.9342			^{188}0.9119	^{187}0.9310			
53	MICROFOCUS-5	^{182}0.9679	^{145}0.9835				^{195}0.9733	^{184}0.8361	^{89}0.8563	^{82}0.8760	^{80}0.8958
54	MICROSOFT-0	^{16}0.0208	^{17}0.0292	^{12}0.0361	^{12}0.0536	^{10}0.1502	^{21}0.0329	^{21}0.0443	^{21}0.0544	^{23}0.0767	^{27}0.1733
55	MICROSOFT-1	^{17}0.0214	^{19}0.0299	^{13}0.0373	^{13}0.0542	^{18}0.1585	^{22}0.0339	^{23}0.0449			
56	MICROSOFT-2	^{25}0.0252	^{26}0.0245	^{25}0.0425	^{16}0.0600	^{12}0.1558	^{35}0.0387	^{34}0.0503			
57	MICROSOFT-3	^{14}0.0123	^{14}0.0193				^{16}0.0223	^{16}0.0304	^{16}0.0384	^{15}0.0570	^{20}0.1602
58	MICROSOFT-4	^{10}0.0128	^{11}0.0179	^{8}0.0241	^{9}0.0405	^{17}0.1628	^{13}0.0209	^{13}0.0288	^{15}0.0360	^{13}0.0550	^{18}0.1576
59	MICROSOFT-5	^{9}0.0119	^{9}0.0171	^{7}0.0218	^{7}0.0387	^{18}0.1654	^{12}0.0201	^{13}0.0279	^{12}0.0347	^{12}0.0545	^{15}0.1549
60	MICROSOFT-6	^{5}0.0058	^{5}0.0080	^{5}0.0110	^{6}0.0284	^{19}0.1664	^{5}0.0109	^{5}0.0141	^{5}0.0183	^{3}0.0363	^{13}0.1544
61	NEC-0	^{60}0.0483	^{57}0.0604	^{37}0.0726	^{35}0.0989	^{36}0.2378	^{64}0.0662	^{63}0.0815	^{47}0.0961	^{44}0.1199	^{37}0.1994
62	NEC-1	^{73}0.0711	^{73}0.0899				^{79}0.0889	^{79}0.1081	^{60}0.1276	^{56}0.1565	^{50}0.2311
63	NEC-2	^{2}0.0019	^{2}0.0024	^{2}0.0038	^{4}0.0211	^{2}0.0991	^{1}0.0040	^{2}0.0047	^{2}0.0057	^{2}0.0090	^{2}0.0723
64	NEC-3	^{1}0.0018	^{1}0.0021	^{1}0.0026	^{1}0.0113	^{1}0.0788	^{2}0.0040	^{1}0.0044	^{1}0.0049	^{1}0.0095	^{1}0.0580
65	NEUROTECHNOLOGY-3	^{159}0.5809	^{134}0.6390				^{172}0.5959	^{172}0.6649	^{88}0.7217	^{80}0.7352	^{78}0.8336
66	NEUROTECHNOLOGY-4	^{55}0.0427	^{55}0.0575	^{36}0.0711	^{34}0.0954	^{24}0.1845	^{55}0.0495	^{54}0.0656	^{44}0.0810	^{39}0.1167	^{42}0.2138
67	NEUROTECHNOLOGY-5	^{50}0.0384	^{51}0.0527	^{30}0.0546	^{27}0.0811	^{7}0.1366	^{42}0.0422	^{48}0.0564	^{37}0.0705	^{36}0.0988	^{35}0.2014
68	NEWLAND-2						^{150}0.4015	^{150}0.4405	^{81}0.4719	^{75}0.5133	
69	NOBLIS-2	^{185}0.9943	^{147}0.9959				^{199}0.9965	^{196}0.9974	^{88}0.9980	^{82}0.9986	
70	NTECHLAB-0	^{65}0.0518	^{63}0.0666	^{40}0.0850	^{38}0.1158		^{67}0.0677	^{64}0.0830	^{50}0.1029	^{45}0.1306	^{34}0.1948
71	NTECHLAB-1	^{67}0.0634	^{70}0.0818	^{45}0.1006	^{42}0.1337	^{33}0.2162	^{75}0.0805	^{76}0.1021			
72	NTECHLAB-3	^{45}0.0329	^{44}0.0434				^{46}0.0445	^{47}0.0561	^{36}0.0699	^{30}0.0933	^{21}0.1609

Table 10: **Identification-mode: Effect of N on FNIR at high threshold.** *Values are threshold-based miss rates i.e.* FNIR *at* FPIR *= 0.001 for five enrollment population sizes, N. The left six columns apply for enrollment of a variable number of images per subject. The right six columns apply for enrollment of one image. Missing entries usually apply because another algorithm from the same developer was run instead. Some developers are missing because less accurate algorithms were not run on galleries with $N \geq 3\,000\,000$. Throughout blue superscripts indicate the rank of the algorithm for that column.*

MISSES BELOW THRESHOLD, T FNIR(N, T > 0, R > L)		ENROL LIFETIME DATASET: FRVT 2018					ENROL MOST RECENT DATASET: FRVT 2018				
#	ALGORITHM	N=0.64M	N=1.6M	N=3.0M	N=6.0M	N=12.0M	N=0.64M	N=1.6M	N=3.0M	N=6.0M	N=12.0M
73	NTECHLAB-4	^{27}0.0253	^{36}0.0337	^{20}0.0433	^{24}0.0692	^{28}0.1845	^{21}0.0337	^{20}0.0431	^{22}0.0545	^{20}0.0749	^{12}0.1528
74	NTECHLAB-5	^{35}0.0268	^{31}0.0347				^{26}0.0358	^{22}0.0448	^{24}0.0561	^{23}0.0785	^{17}0.1572
75	NTECHLAB-6	^{21}0.0227	^{21}0.0301	^{16}0.0395	^{21}0.0654	^{27}0.1897	^{18}0.0311	^{18}0.0391	^{19}0.0496	^{19}0.0696	^{14}0.1548
76	QUANTASOFT-1	^{184}0.9915	^{156}0.9915				^{173}0.6399	^{170}0.6399	^{83}0.6399		^{78}0.6399
77	RANKONE-0	^{106}0.1485	^{103}0.1788	^{65}0.2210	^{64}0.3260	^{43}0.4758	^{116}0.1899	^{115}0.2192	^{78}0.2635	^{79}0.2992	^{73}0.4201
78	RANKONE-1	^{102}0.1211	^{105}0.1549	^{63}0.1804	^{61}0.2371	^{42}0.3530	^{107}0.1542	^{100}0.1683			
79	RANKONE-2	^{77}0.0744	^{76}0.0943				^{92}0.0998	^{85}0.1200	^{64}0.1382	^{61}0.1744	^{62}0.2636
80	RANKONE-3	^{76}0.0744	^{75}0.0943	^{46}0.1120	^{45}0.1490	^{40}0.2946	^{91}0.0998	^{84}0.1200	^{63}0.1382	^{60}0.1744	^{61}0.2636
81	RANKONE-4	^{116}0.1265	^{110}0.1545				^{109}0.1631	^{109}0.1951	^{75}0.2211		
82	RANKONE-5	^{48}0.0347	^{46}0.0447	^{32}0.0571	^{30}0.0847	^{37}0.2549	^{56}0.0499	^{50}0.0617	^{38}0.0728	^{35}0.0984	^{41}0.2031
83	REALNETWORKS-0	^{119}0.2098	^{113}0.2476	^{68}0.2837			^{120}0.2008	^{119}0.2362			
84	REALNETWORKS-2	^{110}0.1688	^{106}0.2049				^{118}0.1974	^{117}0.2341	^{79}0.2691	^{75}0.3186	^{69}0.3261
85	REMARKAI-2	^{74}0.0731	^{79}0.0991				^{85}0.0971	^{91}0.1264	^{68}0.1495	^{65}0.1928	
86	SENSETIME-0	^{8}0.0118	^{8}0.0165				^{111}0.0184	^{9}0.0234	^{9}0.0296	^{9}0.0427	^{8}0.1287
87	SENSETIME-1	^{11}0.0129	^{10}0.0175				^{11}0.0186	^{11}0.0245	^{11}0.0304	^{11}0.0448	^{9}0.1344
88	SHAMAN-3	^{148}0.3506	^{129}0.3921	^{65}0.4295			^{151}0.4179	^{151}0.4527			
89	SHAMAN-7	^{85}0.0924	^{88}0.1112				^{78}0.1256	^{97}0.1436	^{69}0.1610	^{64}0.1901	^{55}0.2480
90	SIAT-1	^{132}0.2695	^{116}0.2727	^{67}0.2758			^{7}0.0160	^{6}0.0201	^{7}0.0260	^{8}0.0380	^{3}0.1069
91	SIAT-2	^{128}0.2198	^{108}0.2239				^{9}0.0179	^{10}0.0242	^{10}0.0301	^{10}0.0434	^{10}0.1377
92	SMILART-4	^{172}0.8381	^{146}0.9569				^{192}0.9260	^{181}0.9683	^{87}0.9913		
93	SYNESIS-3	^{154}0.4748	^{132}0.5296				^{164}0.5353	^{164}0.5832	^{82}0.6123	^{79}0.6489	^{77}0.6838
94	TEVIAN-4	^{72}0.0685	^{72}0.0878	^{45}0.1032			^{83}0.0952	^{86}0.1201			
95	TEVIAN-5	^{66}0.0518	^{64}0.0667				^{72}0.0717	^{68}0.0898	^{54}0.1094	^{48}0.1338	^{30}0.1873
96	TIGER-0	^{136}0.2859	^{123}0.3361	^{60}0.3659	^{57}0.4139		^{143}0.3452	^{143}0.3921			
97	TIGER-2	^{69}0.0511	^{66}0.0698				^{66}0.0671	^{56}0.0888	^{50}0.1065	^{49}0.1361	^{65}0.2284
98	TONGYITRANS-1	^{71}0.0659	^{71}0.0835	^{44}0.1017	^{41}0.1328		^{57}0.0545	^{55}0.0693			
99	TOSHIBA-0	^{40}0.0374	^{57}0.0529				^{54}0.0488	^{53}0.0648	^{43}0.0809	^{39}0.1170	^{43}0.2140
100	VD-0	^{176}0.8686	^{147}0.9048	^{69}0.9242	^{63}0.9381		^{186}0.8892	^{186}0.9171			
101	VD-1	^{108}0.1312	^{102}0.1654				^{110}0.1664	^{113}0.2036	^{77}0.2372	^{72}0.2759	^{70}0.3814
102	VIGILANTSOLUTIONS-3	^{139}0.3061	^{126}0.3568	^{63}0.3861	^{55}0.3861		^{149}0.3648	^{147}0.4097			
103	VISIONLABS-3	^{31}0.0260	^{30}0.0347	^{23}0.0444	^{23}0.0678		^{39}0.0394	^{35}0.0506	^{28}0.0629	^{29}0.0902	
104	VISIONLABS-4	^{40}0.0294	^{40}0.0402				^{50}0.0452	^{45}0.0604	^{39}0.0733	^{34}0.0982	^{31}0.1893
105	VISIONLABS-5	^{28}0.0250	^{37}0.0353	^{22}0.0441	^{19}0.0628	^{21}0.1727	^{40}0.0396	^{39}0.0531	^{30}0.0654	^{28}0.0878	^{32}0.1894
106	VISIONLABS-6	^{13}0.0131	^{13}0.0185				^{15}0.0211	^{15}0.0289	^{14}0.0359	^{16}0.0571	^{16}0.1672
107	VISIONLABS-7	^{12}0.0131	^{12}0.0185	^{9}0.0242	^{10}0.0412	^{9}0.1496	^{14}0.0211	^{14}0.0289	^{13}0.0359	^{14}0.0569	^{19}0.1576
108	VOCORD-3	^{80}0.0969	^{81}0.1295	^{52}0.1627	^{50}0.2361		^{88}0.0973	^{90}0.1258			
109	VOCORD-5	^{75}0.0735	^{84}0.1076				^{99}0.1261	^{102}0.1697	^{76}0.2327	^{76}0.3286	^{75}0.4628
110	YISHENG-1	^{130}0.2539	^{119}0.3002	^{69}0.3366	^{56}0.3892		^{138}0.3026	^{136}0.3483			
111	YITU-0	^{37}0.0279	^{33}0.0358	^{25}0.0468	^{20}0.0636	^{8}0.1389	^{36}0.0388	^{33}0.0502	^{27}0.0622	^{27}0.0862	^{23}0.1621
112	YITU-1	^{32}0.0261	^{28}0.0341	^{21}0.0434	^{17}0.0611	^{6}0.1361	^{29}0.0366	^{29}0.0472			
113	YITU-2	^{6}0.0096	^{6}0.0138	^{6}0.0174	^{5}0.0274	^{5}0.1180	^{6}0.0156	^{7}0.0204	^{6}0.0258	^{7}0.0382	^{6}0.1241
114	YITU-3	^{7}0.0103	^{7}0.0139				^{8}0.0165	^{8}0.0213	^{8}0.0266	^{9}0.0389	^{7}0.1248
115	YITU-4	^{3}0.0052	^{3}0.0074	^{3}0.0097	^{2}0.0187	^{4}0.1153	^{3}0.0093	^{3}0.0123	^{3}0.0159	^{2}0.0273	^{6}0.1107
116	YITU-5	^{4}0.0057	^{4}0.0076	^{4}0.0100	^{3}0.0188	^{3}0.1111	^{4}0.0101	^{4}0.0128	^{4}0.0163	^{3}0.0294	^{5}0.1118

Table 11: **Identification-mode: Effect of N on FNIR at high threshold.** *Values are threshold-based miss rates i.e.* FNIR *at* FPIR *= 0.001 for five enrollment population sizes, N. The left six columns apply for enrollment of a variable number of images per subject. The right six columns apply for enrollment of one image. Missing entries usually apply because another algorithm from the same developer was run instead. Some developers are missing because less accurate algorithms were not run on galleries with $N \geq 3\,000\,000$. Throughout blue superscripts indicate the rank of the algorithm for that column.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

MISSES NOT AT RANK 1		ENROL LIFETIME						ENROL MOST RECENT					
FNIR(N, T = 0, R = 1)		DATASET: FRVT 2018						DATASET: FRVT 2018					
#	ALGORITHM	N=0.64M	N=1.6M	N=3.0M	N=6.0M	N=12.0M	aN^b	N=0.64M	N=1.6M	N=3.0M	N=6.0M	N=12.0M	aN^b
1	3DIVI-3	^{130}0.0494	^{123}0.0645	^{91}0.0759	^{87}0.0898		^{29}0.0014 $N^{0.267\ 91}$	^{162}0.0680	^{152}0.0957				^{53}0.0023 $N^{0.262\ 88}$
2	3DIVI-5	^{85}0.0100	^{86}0.0133				^{16}0.0002 $N^{0.310\ 89}$	^{99}0.0163	^{97}0.0202	^{66}0.0236	^{64}0.0279	^{62}0.0327	^{13}0.0007 $N^{0.230\ 80}$
3	ALCHERA-0	^{91}0.0106	^{83}0.0121	^{47}0.0135	^{42}0.0170		^{76}0.0006 $N^{0.207\ 34}$	^{100}0.0167	^{92}0.0186				^{95}0.0035 $N^{0.117\ 23}$
4	ALCHERA-3	^{95}0.0119	^{91}0.0159				^{80}0.0002 $N^{0.312\ 90}$	^{66}0.0101	^{72}0.0127	^{56}0.0146	^{56}0.0171	^{56}0.0204	^{31}0.0004 $N^{0.236\ 85}$
5	ANKE-0	^{71}0.0077	^{72}0.0100				^{48}0.0002 $N^{0.287\ 79}$	^{87}0.0128	^{88}0.0158	^{63}0.0181	^{61}0.0214	^{59}0.0251	^{36}0.0006 $N^{0.231\ 82}$
6	AWARE-3	^{108}0.0165	^{101}0.0209	^{82}0.0247	^{80}0.0297		^{70}0.0005 $N^{0.263\ 53}$	^{119}0.0264	^{118}0.0332	^{75}0.0387	^{73}0.0456	^{73}0.0532	^{69}0.0011 $N^{0.236\ 70}$
7	AWARE-5	^{107}0.0163	^{100}0.0208				^{58}0.0004 $N^{0.270\ 67}$	^{121}0.0271	^{117}0.0337	^{76}0.0392	^{74}0.0460	^{66}0.0358	^{103}0.0070 $N^{0.109\ 21}$
8	AYONIX-0	^{179}0.4198	^{184}0.4649	^{98}0.4969	^{62}0.5318		^{118}0.1021 $N^{0.106\ 13}$	^{193}0.4095	^{191}0.4519				^{118}0.0973 $N^{0.108\ 20}$
9	AYONIX-2	^{172}0.2192	^{137}0.2606				^{110}0.0176 $N^{0.189\ 30}$	^{187}0.2954	^{188}0.3432	^{86}0.3753	^{82}0.4116	^{79}0.4480	^{111}0.0449 $N^{0.142\ 30}$
10	CAMVI-3	^{98}0.0144	^{117}0.0368	^{67}0.0528	^{60}0.1791		^{3}0.0000 $N^{1.056\ 130}$	^{114}0.0224	^{140}0.0544				^{2}0.0000 $N^{0.969\ 115}$
11	CAMVI-4	^{45}0.0082	^{110}0.0326				^{1}0.0000 $N^{1.300\ 131}$	^{71}0.0146	^{137}0.0490	^{81}0.0741	^{81}0.2582	^{78}0.2586	^{1}0.0000 $N^{1.087\ 116}$
12	COGENT-0	^{88}0.0103	^{77}0.0106	^{41}0.0109	^{36}0.0114	^{31}0.0122	^{94}0.0047 $N^{0.057\ 6}$	^{82}0.0127	^{74}0.0131	^{52}0.0136	^{46}0.0141	^{44}0.0151	^{101}0.0058 $N^{0.058\ 5}$
13	COGENT-1	^{87}0.0103	^{76}0.0106				^{98}0.0074 $N^{0.029\ 4}$	^{85}0.0127	^{75}0.0131	^{51}0.0136	^{45}0.0141	^{43}0.0151	^{100}0.0058 $N^{0.058\ 4}$
14	COGENT-2	^{20}0.0022	^{20}0.0027	^{14}0.0032	^{12}0.0037	^{11}0.0043	^{30}0.0001 $N^{0.232\ 51}$	^{27}0.0054	^{26}0.0062	^{22}0.0067	^{20}0.0075	^{19}0.0085	^{55}0.0007 $N^{0.187\ 33}$
15	COGENT-3	^{31}0.0032	^{29}0.0037	^{17}0.0042	^{16}0.0048	^{15}0.0056	^{111}13.4494 $N^{0.467\ 1}$	^{29}0.0057	^{27}0.0064	^{24}0.0069	^{22}0.0077	^{20}0.0087	^{60}0.0008 $N^{0.144\ 31}$
16	COGNITEC-0	^{92}0.0146	^{96}0.0189				^{29}0.0001 $N^{0.356\ 106}$	^{112}0.0221	^{112}0.0278	^{72}0.0323	^{71}0.0378	^{69}0.0443	^{66}0.0010 $N^{0.237\ 83}$
17	COGNITEC-1	^{63}0.0069	^{66}0.0089	^{40}0.0106	^{38}0.0128	^{36}0.0154	^{48}0.0002 $N^{0.275\ 70}$	^{78}0.0116	^{82}0.0143	^{61}0.0165	^{59}0.0192	^{57}0.0225	^{45}0.0006 $N^{0.226\ 79}$
18	COGNITEC-2	^{28}0.0035	^{34}0.0044	^{24}0.0052	^{22}0.0061	^{22}0.0075	^{36}0.0001 $N^{0.254\ 57}$	^{46}0.0074	^{42}0.0083	^{34}0.0093	^{32}0.0105	^{31}0.0121	^{57}0.0008 $N^{0.166\ 44}$
19	COGNITEC-3	^{42}0.0040	^{39}0.0048	^{30}0.0055	^{25}0.0064	^{23}0.0078	^{61}0.0003 $N^{0.190\ 31}$	^{48}0.0078	^{46}0.0088	^{37}0.0098	^{35}0.0111	^{35}0.0126	^{62}0.0009 $N^{0.164\ 41}$
20	DAHUA-1	^{40}0.0040	^{40}0.0049				^{44}0.0002 $N^{0.242\ 55}$	^{45}0.0074	^{47}0.0089	^{38}0.0102	^{38}0.0115	^{37}0.0135	^{38}0.0005 $N^{0.200\ 62}$
21	DERMALOG-4	^{163}0.0759	^{126}0.0961	^{94}0.1105	^{59}0.1260		^{39}0.0037 $N^{0.227\ 49}$	^{157}0.1040	^{139}0.1274				^{97}0.0054 $N^{0.221\ 76}$
22	DERMALOG-5	^{74}0.0081	^{75}0.0113				^{26}0.0001 $N^{0.353\ 104}$	^{90}0.0135	^{87}0.0171	^{64}0.0223	^{69}0.0312	^{70}0.0470	^{30}0.0004 $N^{0.280\ 100}$
23	DERMALOG-6	^{55}0.0055	^{48}0.0060				^{90}0.0015 $N^{0.095\ 9}$	^{68}0.0095	^{56}0.0102	^{42}0.0107	^{37}0.0115	^{33}0.0125	^{92}0.0027 $N^{0.092\ 14}$
24	EVERAI-0	^{61}0.0065	^{92}0.0166				^{2}0.0000 $N^{1.029\ 109}$	^{59}0.0102	^{99}0.0209	^{74}0.0348			^{3}0.0000 $N^{0.798\ 114}$
25	EVERAI-1	^{21}0.0022	^{21}0.0027				^{37}0.0001 $N^{0.222\ 48}$	^{20}0.0047	^{20}0.0056	^{20}0.0061			^{42}0.0005 $N^{0.166\ 42}$
26	EVERAI-3	^{16}0.0020	^{16}0.0023	^{11}0.0026	^{11}0.0028		^{68}0.0004 $N^{0.113\ 14}$	^{14}0.0041	^{15}0.0047	^{13}0.0052	^{17}0.0059	^{16}0.0066	^{36}0.0005 $N^{0.160\ 39}$
27	EYEDEA-3	^{129}0.0480	^{121}0.0613	^{80}0.0717	^{54}0.0831		^{71}0.0018 $N^{0.245\ 56}$	^{150}0.0663	^{131}0.0824				^{93}0.0028 $N^{0.238\ 87}$
28	GLORY-1	^{139}0.0818	^{125}0.0932	^{82}0.1007	^{56}0.1091		^{102}0.0147 $N^{0.129\ 16}$	^{152}0.1154	^{137}0.1291				^{109}0.0223 $N^{0.123\ 26}$
29	GORILLA-2	^{90}0.0102	^{87}0.0137				^{47}0.0001 $N^{0.321\ 98}$	^{101}0.0170	^{104}0.0220	^{70}0.0261	^{68}0.0311	^{67}0.0376	^{35}0.0005 $N^{0.269\ 105}$
30	HIK-2	^{104}0.0155	^{94}0.0185	^{50}0.0208	^{40}0.0240	^{42}0.0272	^{88}0.0012 $N^{0.193\ 34}$	^{92}0.0147	^{90}0.0172				^{78}0.0015 $N^{0.179\ 47}$
31	HIK-3	^{77}0.0085	^{78}0.0107				^{54}0.0003 $N^{0.255\ 97}$	^{77}0.0115	^{82}0.0141	^{60}0.0164	^{60}0.0194	^{58}0.0228	^{39}0.0005 $N^{0.235\ 74}$
32	HIK-4	^{76}0.0083	^{73}0.0104	^{48}0.0121	^{41}0.0146	^{38}0.0177	^{54}0.0003 $N^{0.260\ 62}$	^{75}0.0112	^{80}0.0138	^{59}0.0159	^{58}0.0188	^{56}0.0220	^{41}0.0005 $N^{0.230\ 67}$
33	HIK-5	^{26}0.0026	^{25}0.0034	^{19}0.0040	^{17}0.0049		^{51}0.0002 $N^{0.199\ 35}$	^{30}0.0057	^{29}0.0067	^{28}0.0075	^{26}0.0087	^{26}0.0103	^{28}0.0004 $N^{0.202\ 60}$
34	IDEMIA-0	^{38}0.0048	^{52}0.0063	^{31}0.0076	^{29}0.0096	^{27}0.0116	^{27}0.0001 $N^{0.304\ 81}$	^{61}0.0093	^{61}0.0113	^{49}0.0131	^{49}0.0153	^{49}0.0182	^{32}0.0004 $N^{0.227\ 80}$
35	IDEMIA-1	^{51}0.0049	^{53}0.0065	^{30}0.0080	^{31}0.0100	^{30}0.0124	^{22}0.0001 $N^{0.320\ 97}$	^{64}0.0096	^{66}0.0116	^{50}0.0135	^{53}0.0162	^{53}0.0194	^{27}0.0004 $N^{0.243\ 94}$
36	IDEMIA-2	^{70}0.0075	^{70}0.0099	^{43}0.0119	^{42}0.0149	^{39}0.0183	^{30}0.0001 $N^{0.304\ 86}$	^{72}0.0105	^{71}0.0126				^{58}0.0008 $N^{0.194\ 59}$
37	IDEMIA-3	^{34}0.0041	^{45}0.0054				^{36}0.0001 $N^{0.294\ 82}$	^{50}0.0080	^{54}0.0095	^{43}0.0110	^{43}0.0127	^{41}0.0148	^{34}0.0005 $N^{0.212\ 70}$
38	IDEMIA-4	^{35}0.0042	^{43}0.0052	^{27}0.0061	^{28}0.0074	^{25}0.0088	^{40}0.0001 $N^{0.287\ 80}$	^{51}0.0080	^{50}0.0092	^{41}0.0106	^{41}0.0124	^{40}0.0143	^{43}0.0005 $N^{0.202\ 61}$
39	IDEMIA-5	^{47}0.0047	^{50}0.0062	^{29}0.0073	^{21}0.0089	^{28}0.0107	^{36}0.0001 $N^{0.280\ 72}$	^{55}0.0090	^{56}0.0107	^{47}0.0123	^{47}0.0144	^{47}0.0169	^{37}0.0005 $N^{0.217\ 74}$
40	IDEMIA-6	^{56}0.0055	^{57}0.0071	^{34}0.0083	^{32}0.0100	^{34}0.0119	^{43}0.0001 $N^{0.270\ 66}$	^{66}0.0102	^{69}0.0122	^{55}0.0139	^{52}0.0161	^{52}0.0187	^{40}0.0006 $N^{0.206\ 69}$
41	IMAGUS-2	^{162}0.1470	^{133}0.1833	^{96}0.2086	^{61}0.2379		^{96}0.0053 $N^{0.215\ 45}$	^{179}0.1838	^{177}0.2223				^{106}0.0115 $N^{0.206\ 67}$
42	INCODE-1	^{83}0.0098	^{84}0.0131	^{84}0.0286	^{53}0.0466		^{4}0.0000 $N^{0.729\ 108}$	^{95}0.0151	^{93}0.0190				^{41}0.0005 $N^{0.230\ 56}$
43	INCODE-3	^{62}0.0067	^{67}0.0088				^{33}0.0001 $N^{0.308\ 88}$	^{80}0.0121	^{86}0.0153	^{62}0.0178	^{62}0.0215	^{60}0.0258	^{29}0.0004 $N^{0.257\ 99}$
44	INNOVATRICS-4	^{65}0.0070	^{61}0.0081				^{81}0.0008 $N^{0.162\ 21}$	^{79}0.0120	^{80}0.0149	^{58}0.0158	^{50}0.0158	^{49}0.0158	^{96}0.0040 $N^{0.088\ 12}$
45	ISYSTEMS-0	^{68}0.0074	^{64}0.0085	^{38}0.0095	^{36}0.0105	^{37}0.0118	^{82}0.0009 $N^{0.160\ 20}$	^{82}0.0122	^{77}0.0136				^{90}0.0025 $N^{0.119\ 25}$
46	ISYSTEMS-1	^{69}0.0074	^{63}0.0085	^{37}0.0094	^{38}0.0105	^{36}0.0118	^{83}0.0009 $N^{0.158\ 15}$	^{81}0.0122	^{76}0.0136				^{91}0.0025 $N^{0.118\ 24}$
47	ISYSTEMS-2	^{39}0.0039	^{37}0.0046	^{23}0.0052			^{66}0.0004 $N^{0.175\ 26}$	^{45}0.0076	^{44}0.0088	^{36}0.0096	^{34}0.0108	^{32}0.0121	^{64}0.0009 $N^{0.156\ 35}$
48	ISYSTEMS-3	^{33}0.0035	^{32}0.0040	^{20}0.0044	^{15}0.0050	^{16}0.0057	^{65}0.0004 $N^{0.166\ 23}$	^{40}0.0069	^{37}0.0075	^{30}0.0081	^{27}0.0090	^{25}0.0100	^{76}0.0012 $N^{0.129\ 28}$
49	LOOKMAN-3	^{78}0.0086	^{67}0.0089				^{98}0.0049 $N^{0.042\ 3}$	^{74}0.0109	^{62}0.0114	^{45}0.0117	^{40}0.0123	^{36}0.0131	^{97}0.0049 $N^{0.059\ 8}$
50	MEGVII-0	^{67}0.0072	^{71}0.0099	^{45}0.0126	^{45}0.0150	^{38}0.0182	^{32}0.0001 $N^{0.317\ 94}$	^{47}0.0075	^{51}0.0094	^{44}0.0111	^{44}0.0134	^{46}0.0162	^{19}0.0002 $N^{0.264\ 103}$
51	MEGVII-1						-	^{84}0.0124	^{78}0.0137	^{57}0.0148	^{54}0.0163	^{50}0.0182	^{87}0.0021 $N^{0.131\ 29}$
52	MICROFOCUS-3	^{181}0.4791	^{146}0.5389	^{99}0.5771			^{107}0.0951 $N^{0.121\ 15}$	^{196}0.5417	^{196}0.5953				^{114}0.1370 $N^{0.103\ 19}$
53	MICROFOCUS-5	^{176}0.3155	^{141}0.3701				^{108}0.0307 $N^{0.174\ 25}$	^{190}0.3716	^{189}0.4257	^{87}0.4624	^{83}0.5013	^{80}0.5404	^{112}0.0684 $N^{0.127\ 27}$
54	MICROSOFT-0	^{18}0.0021	^{19}0.0026	^{13}0.0031	^{14}0.0040	^{13}0.0048	^{16}0.0000 $N^{0.280\ 73}$	^{23}0.0051	^{23}0.0058	^{21}0.0066	^{21}0.0077	^{22}0.0090	^{25}0.0003 $N^{0.198\ 59}$
55	MICROSOFT-1	^{17}0.0020	^{18}0.0026	^{12}0.0031	^{13}0.0038	^{12}0.0047	^{14}0.0000 $N^{0.286\ 78}$	^{21}0.0049	^{21}0.0056				^{47}0.0006 $N^{0.158\ 38}$
56	MICROSOFT-2	^{22}0.0023	^{22}0.0029	^{15}0.0035	^{15}0.0042	^{14}0.0051	^{19}0.0001 $N^{0.272\ 69}$	^{26}0.0052	^{25}0.0061				^{40}0.0005 $N^{0.174\ 48}$
57	MICROSOFT-3	^{2}0.0009	^{4}0.0011				^{12}0.0000 $N^{0.255\ 58}$	^{3}0.0028	^{4}0.0032	^{5}0.0035	^{5}0.0039	^{4}0.0045	^{23}0.0003 $N^{0.166\ 43}$
58	MICROSOFT-4	^{1}0.0008	^{1}0.0010	^{3}0.0013	^{4}0.0015	^{4}0.0019	^{5}0.0000 $N^{0.285\ 76}$	^{2}0.0027	^{2}0.0031	^{4}0.0034	^{3}0.0038	^{3}0.0045	^{17}0.0003 $N^{0.174\ 49}$
59	MICROSOFT-5	^{4}0.0010	^{5}0.0013	^{5}0.0015	^{6}0.0019	^{7}0.0025	^{8}0.0000 $N^{0.304\ 85}$	^{4}0.0028	^{5}0.0033	^{7}0.0037	^{8}0.0044	^{8}0.0052	^{8}0.0002 $N^{0.215\ 72}$
60	MICROSOFT-6	^{5}0.0010	^{7}0.0014	^{7}0.0016	^{9}0.0020	^{9}0.0026	^{6}0.0000 $N^{0.317\ 95}$	^{5}0.0029	^{6}0.0033	^{8}0.0039	^{10}0.0045	^{9}0.0053	^{12}0.0002 $N^{0.206\ 66}$
61	NEC-0	^{82}0.0097	^{83}0.0127	^{48}0.0154	^{46}0.0185	^{40}0.0223	^{53}0.0002 $N^{0.284\ 75}$	^{96}0.0157	^{95}0.0196	^{65}0.0229	^{63}0.0270	^{61}0.0320	^{38}0.0006 $N^{0.248\ 93}$
62	NEC-1	^{97}0.0136	^{97}0.0164				^{85}0.0009 $N^{0.202\ 42}$	^{108}0.0206	^{106}0.0235	^{69}0.0259	^{67}0.0292	^{63}0.0329	^{89}0.0024 $N^{0.160\ 40}$
63	NEC-2	^{6}0.0010	^{3}0.0011	^{1}0.0012	^{1}0.0012	^{1}0.0014	^{67}0.0003 $N^{0.096\ 11}$	^{1}0.0026	^{1}0.0028	^{1}0.0029	^{1}0.0030	^{1}0.0031	^{74}0.0012 $N^{0.059\ 7}$
64	NEC-3	^{7}0.0012	^{6}0.0013	^{4}0.0014	^{2}0.0014	^{2}0.0016	^{73}0.0005 $N^{0.061\ 7}$	^{6}0.0030	^{3}0.0031	^{2}0.0032	^{2}0.0034	^{2}0.0035	^{81}0.0016 $N^{0.048\ 2}$
65	NEUROTECHNOLOGY-3	^{106}0.0161	^{98}0.0199				^{79}0.0007 $N^{0.234\ 52}$	^{107}0.0204	^{109}0.0250	^{71}0.0288	^{70}0.0331	^{68}0.0386	^{70}0.0011 $N^{0.216\ 73}$
66	NEUROTECHNOLOGY-4	^{52}0.0049	^{47}0.0058	^{28}0.0065	^{27}0.0075	^{28}0.0087	^{63}0.0004 $N^{0.195\ 36}$	^{43}0.0072	^{40}0.0082	^{32}0.0090	^{31}0.0100	^{30}0.0114	^{63}0.0009 $N^{0.156\ 34}$
67	NEUROTECHNOLOGY-5	^{34}0.0035	^{33}0.0042	^{19}0.0043	^{20}0.0053	^{17}0.0061	^{59}0.0003 $N^{0.184\ 28}$	^{33}0.0061	^{31}0.0068	^{28}0.0074	^{24}0.0082	^{23}0.0094	^{61}0.0008 $N^{0.149\ 32}$
68	NEWLAND-2						-	^{151}0.0671	^{150}0.0811	^{85}0.0913	^{78}0.1088		^{98}0.0050 $N^{0.195\ 36}$
69	NOBLIS-2	^{157}0.1261	^{132}0.1565				^{96}0.0054 $N^{0.226\ 52}$	^{187}0.1509	^{160}0.1816	^{84}0.2040	^{80}0.2377		^{105}0.0102 $N^{0.201\ 59}$
70	NTECHLAB-0	^{57}0.0056	^{55}0.0077	^{36}0.0094	^{37}0.0114	^{34}0.0139	^{25}0.0001 $N^{0.323\ 99}$	^{60}0.0092	^{63}0.0115	^{54}0.0137	^{55}0.0164	^{54}0.0196	^{17}0.0003 $N^{0.263\ 101}$
71	NTECHLAB-1	^{66}0.0070	^{69}0.0097	^{42}0.0119	^{40}0.0146	^{37}0.0179	^{31}0.0001 $N^{0.317\ 96}$	^{73}0.0108	^{81}0.0139				^{16}0.0003 $N^{0.278\ 108}$
72	NTECHLAB-3	^{37}0.0037	^{41}0.0051				^{13}0.0000 $N^{0.351\ 103}$	^{35}0.0065	^{41}0.0082	^{35}0.0096	^{36}0.0115	^{38}0.0135	^{15}0.0002 $N^{0.251\ 97}$

Table 12: **Investigation-mode: Effect of N on FNIR at rank 1** *For five enrollment population sizes, N, with T = 0 and FPIR = 1. The left five columns apply for consolidated enrollment of a variable number of lifetime images from each subject. The right five columns apply for enrollment of one recent image. Missing entries usually apply because another algorithm from the same developer was run instead. Some developers are missing because less accurate algorithms were not run on galleries with N > 1 600 000. Throughout blue superscripts indicate the rank of the algorithm for that column, and yellow highlighting indicates the most accurate value. Caution: The Power-low models are mostly intended to draw attention to the kind of behavior, not as a model to be used for prediction.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

MISSES NOT AT RANK 1		ENROL LIFETIME						ENROL MOST RECENT					
FNIR(N, T = 0, R = 1)		DATASET: FRVT 2018						DATASET: FRVT 2018					
#	ALGORITHM	N=0.64M	N=1.6M	N=3.0M	N=6.0M	N=12.0M	aN^b	N=0.64M	N=1.6M	N=3.0M	N=6.0M	N=12.0M	aN^b
73	NTECHLAB-4	[30] 0.0030	[32] 0.0040	[21] 0.0049	[22] 0.0060	[21] 0.0075	[19] 0.0000 $N^{0.315}$ [98]	[38] 0.0056	[32] 0.0068	[29] 0.0078	[28] 0.0092	[28] 0.0107	[20] 0.0005 $N^{0.220}$ [76]
74	NTECHLAB-5	[27] 0.0028	[31] 0.0039				[10] 0.0000 $N^{0.361}$ [105]	[34] 0.0061	[29] 0.0064	[28] 0.0076	[30] 0.0092	[29] 0.0112	[7] 0.0001 $N^{0.286}$ [104]
75	NTECHLAB-6	[28] 0.0024	[26] 0.0034	[18] 0.0042	[19] 0.0052	[18] 0.0066	[11] 0.0000 $N^{0.346}$ [102]	[19] 0.0047	[24] 0.0059	[23] 0.0069	[23] 0.0081	[24] 0.0098	[10] 0.0002 $N^{0.250}$ [95]
76	QUANTASOFT-1	[180] 0.9857	[149] 0.9857				-	[182] 0.2198	[176] 0.2198	[85] 0.2198		[86] 0.2198	[116] 0.2198 $N^{0.000}$ [7]
77	RANKONE-0	[122] 0.0255	[108] 0.0319	[55] 0.0366	[52] 0.0425	[44] 0.0486	[88] 0.0014 $N^{0.232}$ [47]	[137] 0.0375	[133] 0.0455	[80] 0.0514	[75] 0.0564	[79] 0.0654	[94] 0.0032 $N^{0.186}$ [31]
78	RANKONE-1	[102] 0.0152	[97] 0.0194	[57] 0.0224	[49] 0.0260	[43] 0.0302	[77] 0.0007 $N^{0.232}$ [50]	[115] 0.0226	[106] 0.0247				[102] 0.0062 $N^{0.197}$ [16]
79	RANKONE-2	[99] 0.0117	[89] 0.0149				[82] 0.0003 $N^{0.288}$ [85]	[106] 0.0181	[102] 0.0221	[88] 0.0250	[66] 0.0288	[65] 0.0330	[72] 0.0012 $N^{0.204}$ [25]
80	RANKONE-3	[95] 0.0117	[88] 0.0149	[49] 0.0172	[47] 0.0200	[41] 0.0236	[71] 0.0005 $N^{0.237}$ [54]	[105] 0.0181	[101] 0.0221	[87] 0.0250	[65] 0.0288	[64] 0.0330	[71] 0.0012 $N^{0.204}$ [64]
81	RANKONE-4	[123] 0.0246	[107] 0.0318				[74] 0.0006 $N^{0.282}$ [74]	[130] 0.0351	[132] 0.0441	[79] 0.0508			[77] 0.0014 $N^{0.259}$ [93]
82	RANKONE-5	[59] 0.0058	[58] 0.0072	[36] 0.0086	[33] 0.0103	[32] 0.0122	[37] 0.0002 $N^{0.236}$ [60]	[67] 0.0102	[66] 0.0120	[53] 0.0136	[51] 0.0158	[51] 0.0182	[54] 0.0007 $N^{0.231}$ [38]
83	REALNETWORKS-0	[131] 0.0397	[115] 0.0443	[56] 0.0527			[75] 0.0007 $N^{0.290}$ [81]	[127] 0.0380	[131] 0.0426				[56] 0.0008 $N^{0.280}$ [110]
84	REALNETWORKS-2	[117] 0.0240	[103] 0.0320				[64] 0.0004 $N^{0.313}$ [91]	[125] 0.0323	[126] 0.0418	[78] 0.0494	[77] 0.0587	[75] 0.0604	[82] 0.0017 $N^{0.223}$ [77]
85	REMARKAI-2	[46] 0.0047	[51] 0.0062				[23] 0.0001 $N^{0.314}$ [92]	[56] 0.0085	[58] 0.0105	[46] 0.0122	[48] 0.0145		[36] 0.0004 $N^{0.237}$ [86]
86	SENSETIME-0	[13] 0.0016	[13] 0.0018				-	[17] 0.0046	[16] 0.0048	[17] 0.0050	[14] 0.0053	[13] 0.0057	[84] 0.0018 $N^{0.071}$ [9]
87	SENSETIME-1	[12] 0.0016	[11] 0.0018				-	[16] 0.0046	[17] 0.0048	[15] 0.0050	[15] 0.0053	[14] 0.0062	[76] 0.0012 $N^{0.085}$ [15]
88	SHAMAN-3	[148] 0.0908	[127] 0.0969	[63] 0.1091			[97] 0.0060 $N^{0.196}$ [37]	[158] 0.1074	[155] 0.1266				[104] 0.0097 $N^{0.180}$ [80]
89	SHAMAN-7	[126] 0.0290	[105] 0.0310				[101] 0.0106 $N^{0.075}$ [8]	[139] 0.0397	[129] 0.0422	[77] 0.0442	[75] 0.0468	[72] 0.0499	[107] 0.0139 $N^{0.075}$ [10]
90	SIAT-1	[174] 0.2638	[138] 0.2639	[66] 0.2640			[110] 0.2618 $N^{0.001}$ [3]	[11] 0.0037	[10] 0.0039	[10] 0.0041	[9] 0.0044	[6] 0.0049	[85] 0.0010 $N^{0.095}$ [17]
91	SIAT-2	[171] 0.2127	[136] 0.2128				[109] 0.2115 $N^{0.000}$ [2]	[12] 0.0037	[11] 0.0040	[11] 0.0042	[11] 0.0045	[5] 0.0049	[87] 0.0011 $N^{0.092}$ [13]
92	SMILART-4	[182] 0.8189	[147] 0.9531				[106] 0.0894 $N^{0.166}$ [22]	[195] 0.9176	[196] 0.9649	[88] 0.9908			[118] 0.4706 $N^{0.050}$ [3]
93	SYNESIS-3	[154] 0.1133	[131] 0.1350				[100] 0.0088 $N^{0.191}$ [32]	[166] 0.1478	[167] 0.1721	[83] 0.1897	[77] 0.2108	[77] 0.2338	[108] 0.0184 $N^{0.156}$ [36]
94	TEVIAN-4	[58] 0.0058	[60] 0.0080	[37] 0.0097			[18] 0.0001 $N^{0.381}$ [101]	[71] 0.0105	[75] 0.0134				[28] 0.0003 $N^{0.364}$ [102]
95	TEVIAN-5	[42] 0.0040	[44] 0.0053				[21] 0.0001 $N^{0.307}$ [87]	[44] 0.0074	[48] 0.0092	[39] 0.0104	[42] 0.0125	[42] 0.0151	[22] 0.0003 $N^{0.240}$ [72]
96	TIGER-0	[134] 0.0364	[117] 0.0480	[58] 0.0565	[55] 0.0678		[84] 0.0009 $N^{0.276}$ [71]	[143] 0.0494	[144] 0.0638				[74] 0.0012 $N^{0.279}$ [109]
97	TIGER-2	[32] 0.0034	[28] 0.0044				[20] 0.0001 $N^{0.296}$ [83]	[38] 0.0063	[39] 0.0075	[32] 0.0088	[33] 0.0107	[34] 0.0126	[16] 0.0003 $N^{0.239}$ [88]
98	TONGYITRANS-1	[81] 0.0096	[80] 0.0114	[46] 0.0127	[42] 0.0148		[80] 0.0007 $N^{0.198}$ [33]	[57] 0.0080	[52] 0.0095				[90] 0.0006 $N^{0.187}$ [54]
99	TOSHIBA-0	[25] 0.0026	[24] 0.0033				[17] 0.0001 $N^{0.285}$ [77]	[32] 0.0058	[32] 0.0068	[27] 0.0076	[25] 0.0085	[46] 0.0178	[5] 0.0001 $N^{0.337}$ [112]
100	VD-0	[160] 0.3583	[143] 0.4303	[67] 0.4776	[62] 0.5281		[105] 0.0355 $N^{0.179}$ [24]	[192] 0.4073	[192] 0.4751				[110] 0.0431 $N^{0.180}$ [56]
101	VD-1	[113] 0.0184	[100] 0.0221				[87] 0.0012 $N^{0.201}$ [41]	[122] 0.0256	[115] 0.0302	[73] 0.0341	[72] 0.0389	[70] 0.0443	[85] 0.0021 $N^{0.198}$ [53]
102	VIGILANTSOLUTIONS-3	[138] 0.0410	[127] 0.0549	[59] 0.0654	[54] 0.0654		[92] 0.0023 $N^{0.219}$ [46]	[148] 0.0561	[148] 0.0719				[79] 0.0015 $N^{0.271}$ [107]
103	VISIONLABS-3	[38] 0.0037	[41] 0.0050	[30] 0.0076	[39] 0.0130		[5] 0.0000 $N^{0.563}$ [107]	[42] 0.0070	[48] 0.0089	[48] 0.0124	[57] 0.0185		[3] 0.0000 $N^{0.491}$ [113]
104	VISIONLABS-4	[14] 0.0016	[14] 0.0020				[34] 0.0001 $N^{0.203}$ [43]	[13] 0.0037	[13] 0.0044	[14] 0.0049	[19] 0.0062	[23] 0.0088	[6] 0.0001 $N^{0.262}$ [111]
105	VISIONLABS-5	[11] 0.0015	[12] 0.0018	[9] 0.0020	[10] 0.0028	[10] 0.0040	[7] 0.0000 $N^{0.382}$ [100]	[9] 0.0035	[12] 0.0041	[12] 0.0046	[16] 0.0054	[17] 0.0068	[11] 0.0002 $N^{0.228}$ [78]
106	VISIONLABS-6	[10] 0.0013	[9] 0.0015				[52] 0.0002 $N^{0.142}$ [18]	[7] 0.0030	[7] 0.0033	[6] 0.0037	[7] 0.0044	[12] 0.0057	[9] 0.0002 $N^{0.214}$ [71]
107	VISIONLABS-7	[6] 0.0013	[8] 0.0014	[6] 0.0016	[5] 0.0018	[5] 0.0022	[33] 0.0001 $N^{0.183}$ [27]	[6] 0.0030	[6] 0.0033	[4] 0.0035	[4] 0.0039	[7] 0.0050	[21] 0.0003 $N^{0.165}$ [46]
108	VOCORD-3	[54] 0.0055	[55] 0.0067	[32] 0.0090	[30] 0.0095		[42] 0.0001 $N^{0.271}$ [68]	[41] 0.0070	[43] 0.0085				[33] 0.0005 $N^{0.204}$ [63]
109	VOCORD-5	[48] 0.0049	[46] 0.0057				[57] 0.0004 $N^{0.187}$ [29]	[54] 0.0081	[45] 0.0092	[40] 0.0104	[39] 0.0120	[39] 0.0140	[61] 0.0006 $N^{0.188}$ [59]
110	YISHENG-1	[108] 0.0155	[93] 0.0208	[58] 0.0248	[51] 0.0298		[80] 0.0003 $N^{0.294}$ [81]	[136] 0.0227	[134] 0.0290				[52] 0.0006 $N^{0.258}$ [105]
111	YITU-0	[41] 0.0040	[48] 0.0047	[25] 0.0053	[24] 0.0061	[20] 0.0071	[58] 0.0003 $N^{0.202}$ [40]	[35] 0.0066	[38] 0.0074	[31] 0.0082	[29] 0.0092	[27] 0.0103	[89] 0.0008 $N^{0.156}$ [57]
112	YITU-1	[38] 0.0039	[39] 0.0046	[22] 0.0051	[21] 0.0059	[19] 0.0069	[56] 0.0003 $N^{0.192}$ [35]	[37] 0.0065	[35] 0.0072				[80] 0.0015 $N^{0.110}$ [22]
113	YITU-2	[8] 0.0013	[10] 0.0015	[8] 0.0017	[7] 0.0019	[9] 0.0023	[26] 0.0001 $N^{0.196}$ [38]	[15] 0.0041	[14] 0.0044	[13] 0.0047	[13] 0.0050	[13] 0.0055	[88] 0.0011 $N^{0.099}$ [18]
114	YITU-3	[13] 0.0021	[17] 0.0023				[75] 0.0006 $N^{0.098}$ [12]	[25] 0.0052	[19] 0.0054	[19] 0.0057	[18] 0.0061	[15] 0.0065	[83] 0.0017 $N^{0.081}$ [11]
115	YITU-4	[3] 0.0010	[2] 0.0011	[2] 0.0012	[2] 0.0014	[3] 0.0019	[47] 0.0002 $N^{0.130}$ [17]	[10] 0.0036	[9] 0.0037	[9] 0.0040	[6] 0.0042	[18] 0.0072	[13] 0.0002 $N^{0.206}$ [68]
116	YITU-5	[18] 0.0019	[18] 0.0020	[10] 0.0021	[9] 0.0023	[8] 0.0025	[72] 0.0005 $N^{0.095}$ [10]	[18] 0.0047	[19] 0.0048	[16] 0.0050	[10] 0.0052	[10] 0.0055	[86] 0.0021 $N^{0.038}$ [6]

Table 13: **Investigation-mode: Effect of N on FNIR at rank 1** *For five enrollment population sizes, N, with $T = 0$ and $FPIR = 1$. The left five columns apply for consolidated enrollment of a variable number of lifetime images from each subject. The right five columns apply for enrollment of one recent image. Missing entries usually apply because another algorithm from the same developer was run instead. Some developers are missing because less accurate algorithms were not run on galleries with $N > 1\,600\,000$. Throughout blue superscripts indicate the rank of the algorithm for that column, and yellow highlighting indicates the most accurate value. Caution: The Power-low models are mostly intended to draw attention to the kind of behavior, not as a model to be used for prediction.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

MISSES NOT AT RANK 50		ENROL LIFETIME						ENROL MOST RECENT					
FNIR(N, T = 0, R = 50)		DATASET: FRVT 2018						DATASET: FRVT 2018					
#	ALGORITHM	N=0.64M	N=1.6M	N=3.0M	N=6.0M	N=12.0M	aN^b	N=0.64M	N=1.6M	N=3.0M	N=6.0M	N=12.0M	aN^b
1	3DIVI-3	^{127}0.0103	^{116}0.0151	^{89}0.0192	^{96}0.0241		^{30}0.0001 $N^{0.362}$ ^{100}	^{136}0.0159	^{138}0.0217				^{71}0.0002 $N^{0.303}$ ^{106}
2	3DIVI-5	^{79}0.0030	^{77}0.0037				^{52}0.0001 $N^{0.287}$ ^{87}	^{91}0.0065	^{92}0.0074	^{62}0.0083	^{62}0.0094	^{80}0.0107	^{58}0.0007 $N^{0.180}$ ^{71}
3	ALCHERA-0	^{116}0.0073	^{97}0.0076	^{62}0.0079	^{49}0.0101		^{90}0.0012 $N^{0.133}$ ^{38}	^{128}0.0125	^{121}0.0129				^{109}0.0079 $N^{0.034}$ ^{13}
4	ALCHERA-3	^{78}0.0030	^{90}0.0040				^{22}0.0000 $N^{0.309}$ ^{81}	^{57}0.0047	^{62}0.0052	^{46}0.0056	^{49}0.0063	^{49}0.0070	^{62}0.0008 $N^{0.136}$ ^{80}
5	ANKE-0	^{65}0.0024	^{71}0.0030				^{48}0.0001 $N^{0.234}$ ^{68}	^{83}0.0057	^{81}0.0065	^{66}0.0072	^{57}0.0081	^{54}0.0092	^{39}0.0006 $N^{0.164}$ ^{83}
6	AWARE-3	^{90}0.0039	^{87}0.0050	^{48}0.0061	^{47}0.0077		^{36}0.0001 $N^{0.299}$ ^{81}	^{106}0.0081	^{113}0.0101	^{71}0.0118	^{69}0.0139	^{73}0.0170	^{27}0.0003 $N^{0.258}$ ^{74}
7	AWARE-5	^{94}0.0041	^{92}0.0053				^{51}0.0001 $N^{0.263}$ ^{76}	^{112}0.0088	^{116}0.0108	^{73}0.0127	^{72}0.0154	^{82}0.0115	^{79}0.0017 $N^{0.128}$ ^{54}
8	AYONIX-0	^{177}0.1723	^{143}0.2142	^{97}0.2467	^{69}0.2850		^{105}0.0085 $N^{0.225}$ ^{62}	^{193}0.1967	^{194}0.2402				^{111}0.0107 $N^{0.218}$ ^{57}
9	AYONIX-2	^{148}0.0646	^{133}0.0873				^{35}0.0008 $N^{0.327}$ ^{90}	^{186}0.0974	^{186}0.1298	^{90}0.1547	^{81}0.1850	^{78}0.2171	^{91}0.0026 $N^{0.273}$ ^{99}
10	CAMVI-3	^{134}0.0142	^{136}0.0367	^{63}0.0527	^{61}0.1789		^{4}0.0000 $N^{1.080}$ ^{108}	^{144}0.0221	^{160}0.0541				^{3}0.0000 $N^{0.980}$ ^{115}
11	CAMVI-4	^{132}0.0078	^{124}0.0323				^{1}0.0000 $N^{1.549}$ ^{111}	^{133}0.0137	^{157}0.0485	^{83}0.0736	^{82}0.2380	^{79}0.2383	^{1}0.0000 $N^{1.024}$ ^{116}
12	COGENT-0	^{55}0.0021	^{55}0.0024	^{31}0.0027	^{31}0.0031	^{38}0.0045	^{39}0.0001 $N^{0.253}$ ^{73}	^{59}0.0047	^{54}0.0050	^{41}0.0054	^{48}0.0062	^{63}0.0122	^{10}0.0001 $N^{0.288}$ ^{102}
13	COGENT-1	^{54}0.0021	^{57}0.0024				^{60}0.0002 $N^{0.189}$ ^{54}	^{58}0.0047	^{53}0.0050	^{40}0.0054	^{47}0.0062	^{62}0.0122	^{9}0.0001 $N^{0.288}$ ^{101}
14	COGENT-2	^{24}0.0011	^{27}0.0013	^{17}0.0014	^{17}0.0016	^{14}0.0017	^{68}0.0002 $N^{0.137}$ ^{40}	^{37}0.0038	^{36}0.0041	^{25}0.0042	^{26}0.0044	^{23}0.0047	^{75}0.0016 $N^{0.066}$ ^{32}
15	COGENT-3	^{35}0.0014	^{31}0.0016	^{19}0.0018	^{15}0.0020	^{17}0.0023	^{111}35.4798 $N^{0.878}$ ^{1}	^{38}0.0040	^{40}0.0042	^{32}0.0044	^{30}0.0046	^{26}0.0048	^{77}0.0017 $N^{0.068}$ ^{30}
16	COGNITEC-0	^{92}0.0039	^{87}0.0050				^{6}0.0000 $N^{0.899}$ ^{106}	^{110}0.0076	^{108}0.0092	^{65}0.0104	^{68}0.0123	^{68}0.0148	^{39}0.0004 $N^{0.278}$ ^{66}
17	COGNITEC-1	^{67}0.0024	^{66}0.0028	^{39}0.0032	^{39}0.0037	^{32}0.0044	^{61}0.0002 $N^{0.200}$ ^{55}	^{81}0.0056	^{76}0.0060	^{57}0.0066	^{58}0.0072	^{57}0.0081	^{65}0.0010 $N^{0.128}$ ^{55}
18	COGNITEC-2	^{49}0.0020	^{43}0.0021	^{34}0.0023	^{22}0.0025	^{19}0.0027	^{74}0.0004 $N^{0.113}$ ^{33}	^{68}0.0049	^{63}0.0052	^{42}0.0054	^{38}0.0056	^{35}0.0060	^{86}0.0021 $N^{0.063}$ ^{28}
19	COGNITEC-3	^{61}0.0023	^{54}0.0025	^{29}0.0026	^{28}0.0028	^{21}0.0031	^{81}0.0007 $N^{0.084}$ ^{25}	^{75}0.0053	^{71}0.0056	^{49}0.0057	^{42}0.0060	^{40}0.0063	^{90}0.0025 $N^{0.057}$ ^{24}
20	DAHUA-1	^{53}0.0021	^{58}0.0022				^{78}0.0005 $N^{0.099}$ ^{30}	^{56}0.0046	^{50}0.0049	^{37}0.0051	^{36}0.0054	^{32}0.0058	^{70}0.0015 $N^{0.085}$ ^{40}
21	DERMALOG-4	^{139}0.0186	^{121}0.0272	^{61}0.0340	^{58}0.0427		^{53}0.0001 $N^{0.372}$ ^{96}	^{147}0.0262	^{153}0.0365				^{14}0.0002 $N^{0.363}$ ^{108}
22	DERMALOG-5	^{112}0.0066	^{107}0.0092				^{24}0.0001 $N^{0.362}$ ^{95}	^{125}0.0113	^{126}0.0142	^{78}0.0192	^{76}0.0275	^{75}0.0427	^{37}0.0004 $N^{0.248}$ ^{98}
23	DERMALOG-6	^{100}0.0046	^{84}0.0047				^{100}0.0035 $N^{0.020}$ ^{7}	^{105}0.0080	^{96}0.0081	^{63}0.0083	^{60}0.0085	^{52}0.0087	^{106}0.0053 $N^{0.030}$ ^{10}
24	EVERAI-0	^{104}0.0050	^{113}0.0150				^{3}0.0000 $N^{1.185}$ ^{109}	^{101}0.0077	^{138}0.0182	^{79}0.0317			^{2}0.0000 $N^{0.919}$ ^{114}
25	EVERAI-1	^{30}0.0013	^{29}0.0014				^{65}0.0004 $N^{0.096}$ ^{29}	^{25}0.0031	^{28}0.0033	^{21}0.0034			^{30}0.0012 $N^{0.070}$ ^{35}
26	EVERAI-3	^{28}0.0012	^{28}0.0013	^{16}0.0014	^{14}0.0014		^{79}0.0006 $N^{0.057}$ ^{16}	^{21}0.0029	^{18}0.0030	^{14}0.0032	^{15}0.0034	^{12}0.0036	^{69}0.0012 $N^{0.068}$ ^{31}
27	EYEDEA-3	^{130}0.0113	^{117}0.0160	^{80}0.0209	^{56}0.0252		^{44}0.0001 $N^{0.364}$ ^{96}	^{140}0.0175	^{139}0.0236				^{19}0.0002 $N^{0.336}$ ^{104}
28	GLORY-1	^{138}0.0415	^{125}0.0490	^{66}0.0569	^{59}0.0600		^{102}0.0047 $N^{0.164}$ ^{45}	^{173}0.0604	^{171}0.0698				^{108}0.0073 $N^{0.188}$ ^{85}
29	GORILLA-2	^{57}0.0023	^{76}0.0029				^{20}0.0000 $N^{0.287}$ ^{78}	^{70}0.0050	^{77}0.0061	^{58}0.0070	^{58}0.0084	^{56}0.0102	^{15}0.0002 $N^{0.232}$ ^{92}
30	HIK-2	^{125}0.0084	^{106}0.0090	^{59}0.0097	^{50}0.0106	^{43}0.0118	^{96}0.0018 $N^{0.115}$ ^{34}	^{111}0.0087	^{104}0.0093				^{98}0.0035 $N^{0.092}$ ^{34}
31	HIK-3	^{58}0.0023	^{64}0.0028				^{47}0.0001 $N^{0.230}$ ^{54}	^{46}0.0044	^{47}0.0051	^{50}0.0058	^{51}0.0066	^{47}0.0076	^{34}0.0003 $N^{0.189}$ ^{78}
32	HIK-4	^{59}0.0023	^{69}0.0028	^{40}0.0033	^{36}0.0039	^{35}0.0048	^{42}0.0001 $N^{0.246}$ ^{59}	^{52}0.0045	^{58}0.0051	^{51}0.0058	^{50}0.0065	^{46}0.0076	^{38}0.0004 $N^{0.176}$ ^{73}
33	HIK-5	^{13}0.0009	^{17}0.0011	^{13}0.0012	^{19}0.0014		^{86}0.0001 $N^{0.140}$ ^{42}	^{23}0.0029	^{25}0.0033	^{22}0.0035	^{20}0.0038	^{17}0.0042	^{45}0.0006 $N^{0.122}$ ^{52}
34	IDEMIA-0	^{39}0.0016	^{42}0.0019	^{35}0.0023	^{26}0.0026	^{23}0.0031	^{41}0.0001 $N^{0.226}$ ^{62}	^{53}0.0045	^{55}0.0051	^{43}0.0055	^{44}0.0060	^{44}0.0067	^{61}0.0008 $N^{0.134}$ ^{58}
35	IDEMIA-1	^{46}0.0019	^{51}0.0024	^{38}0.0029	^{36}0.0036	^{39}0.0046	^{17}0.0000 $N^{0.307}$ ^{82}	^{67}0.0049	^{73}0.0058	^{56}0.0065	^{56}0.0076	^{59}0.0089	^{33}0.0003 $N^{0.201}$ ^{63}
36	IDEMIA-2	^{82}0.0031	^{77}0.0040	^{43}0.0048	^{43}0.0058	^{40}0.0074	^{31}0.0001 $N^{0.290}$ ^{79}	^{90}0.0061	^{87}0.0069				^{96}0.0010 $N^{0.138}$ ^{59}
37	IDEMIA-3	^{46}0.0019	^{48}0.0022				^{64}0.0002 $N^{0.176}$ ^{48}	^{64}0.0049	^{65}0.0053	^{47}0.0057	^{46}0.0062	^{43}0.0067	^{47}0.0011 $N^{0.109}$ ^{49}
38	IDEMIA-4	^{37}0.0015	^{36}0.0017	^{21}0.0020	^{21}0.0023	^{20}0.0028	^{46}0.0001 $N^{0.207}$ ^{56}	^{46}0.0043	^{40}0.0046	^{36}0.0051	^{37}0.0055	^{29}0.0062	^{63}0.0008 $N^{0.121}$ ^{61}
39	IDEMIA-5	^{46}0.0018	^{49}0.0023	^{28}0.0026	^{34}0.0033	^{31}0.0042	^{18}0.0000 $N^{0.289}$ ^{77}	^{61}0.0048	^{70}0.0056	^{54}0.0062	^{54}0.0070	^{50}0.0080	^{40}0.0005 $N^{0.175}$ ^{72}
40	IDEMIA-6	^{57}0.0022	^{61}0.0028	^{41}0.0034	^{35}0.0043	^{36}0.0055	^{37}0.0001 $N^{0.258}$ ^{74}	^{76}0.0054	^{75}0.0062	^{60}0.0072	^{59}0.0084	^{57}0.0102	^{25}0.0003 $N^{0.230}$ ^{98}
41	IMAGUS-2	^{110}0.0348	^{130}0.0510	^{65}0.0641	^{60}0.0904		^{67}0.0002 $N^{0.375}$ ^{97}	^{166}0.0468	^{166}0.0657				^{32}0.0003 $N^{0.371}$ ^{109}
42	INCODE-1	^{71}0.0026	^{74}0.0033	^{58}0.0167	^{57}0.0323		^{2}0.0000 $N^{1.217}$ ^{110}	^{78}0.0055	^{79}0.0063				^{55}0.0007 $N^{0.183}$ ^{84}
43	INCODE-3	^{42}0.0017	^{44}0.0021				^{28}0.0001 $N^{0.251}$ ^{70}	^{50}0.0044	^{61}0.0052	^{48}0.0057	^{52}0.0067	^{49}0.0078	^{31}0.0003 $N^{0.194}$ ^{82}
44	INNOVATRICS-4	^{50}0.0020	^{47}0.0022				^{73}0.0004 $N^{0.118}$ ^{35}	^{72}0.0052	^{73}0.0058	^{53}0.0061	^{45}0.0061	^{38}0.0061	^{92}0.0026 $N^{0.034}$ ^{22}
45	ISYSTEMS-0	^{101}0.0048	^{88}0.0050	^{45}0.0053	^{42}0.0056	^{39}0.0060	^{94}0.0017 $N^{0.076}$ ^{21}	^{108}0.0086	^{102}0.0089				^{103}0.0048 $N^{0.044}$ ^{15}
46	ISYSTEMS-1	^{102}0.0048	^{90}0.0050	^{46}0.0053	^{41}0.0056	^{38}0.0060	^{95}0.0017 $N^{0.078}$ ^{20}	^{107}0.0086	^{101}0.0089				^{104}0.0049 $N^{0.040}$ ^{14}
47	ISYSTEMS-2	^{74}0.0026	^{67}0.0027	^{35}0.0029			^{89}0.0012 $N^{0.061}$ ^{17}	^{77}0.0054	^{72}0.0056	^{52}0.0058	^{43}0.0060	^{41}0.0063	^{93}0.0027 $N^{0.051}$ ^{30}
48	ISYSTEMS-3	^{69}0.0025	^{80}0.0026	^{30}0.0027	^{25}0.0028	^{21}0.0030	^{91}0.0012 $N^{0.053}$ ^{13}	^{73}0.0052	^{68}0.0054	^{42}0.0055	^{39}0.0057	^{33}0.0059	^{94}0.0028 $N^{0.046}$ ^{18}
49	LOOKMAN-3	^{118}0.0075	^{100}0.0077				^{103}0.0060 $N^{0.017}$ ^{6}	^{119}0.0099	^{112}0.0100	^{68}0.0101	^{64}0.0102	^{58}0.0104	^{110}0.0079 $N^{0.016}$ ^{4}
50	MEGVII-0	^{37}0.0012	^{40}0.0019	^{36}0.0025	^{30}0.0032	^{30}0.0041	^{7}0.0000 $N^{0.422}$ ^{103}	^{12}0.0026	^{20}0.0031	^{30}0.0034	^{33}0.0039	^{34}0.0048	^{12}0.0002 $N^{0.304}$ ^{84}
51	MEGVII-1							^{116}0.0091	^{106}0.0094	^{64}0.0097	^{63}0.0101	^{59}0.0106	^{102}0.0044 $N^{0.053}$ ^{21}
52	MICROFOCUS-3	^{179}0.2047	^{148}0.2625	^{84}0.3017			^{104}0.0070 $N^{0.382}$ ^{71}	^{196}0.2518	^{196}0.3113				^{112}0.0114 $N^{0.332}$ ^{91}
53	MICROFOCUS-5	^{154}0.1040	^{140}0.1422				^{86}0.0011 $N^{0.341}$ ^{90}	^{190}0.1322	^{190}0.1744	^{87}0.2066	^{83}0.2445	^{80}0.2829	^{101}0.0042 $N^{0.300}$ ^{96}
54	MICROSOFT-0	^{8}0.0008	^{12}0.0010	^{11}0.0011	^{11}0.0012	^{10}0.0014	^{42}0.0001 $N^{0.174}$ ^{47}	^{16}0.0028	^{17}0.0031	^{13}0.0032	^{17}0.0035	^{14}0.0037	^{58}0.0007 $N^{0.101}$ ^{45}
55	MICROSOFT-1	^{9}0.0008	^{10}0.0009	^{10}0.0011	^{10}0.0012	^{11}0.0014	^{35}0.0001 $N^{0.177}$ ^{50}	^{15}0.0028	^{15}0.0030				^{60}0.0007 $N^{0.098}$ ^{43}
56	MICROSOFT-2	^{11}0.0008	^{11}0.0010	^{9}0.0011	^{12}0.0012	^{13}0.0014	^{30}0.0001 $N^{0.186}$ ^{52}	^{20}0.0029	^{21}0.0032				^{59}0.0007 $N^{0.101}$ ^{36}
57	MICROSOFT-3	^{2}0.0004	^{4}0.0004				^{23}0.0001 $N^{0.152}$ ^{43}	^{4}0.0018	^{4}0.0019	^{4}0.0021	^{4}0.0022	^{3}0.0023	^{52}0.0006 $N^{0.078}$ ^{38}
58	MICROSOFT-4	^{1}0.0004	^{1}0.0004	^{1}0.0005	^{1}0.0005	^{1}0.0006	^{27}0.0001 $N^{0.140}$ ^{41}	^{3}0.0018	^{3}0.0019	^{3}0.0020	^{2}0.0021	^{2}0.0022	^{54}0.0007 $N^{0.070}$ ^{36}
59	MICROSOFT-5	^{3}0.0004	^{3}0.0004	^{3}0.0005	^{2}0.0005	^{2}0.0006	^{34}0.0001 $N^{0.134}$ ^{39}	^{2}0.0018	^{1}0.0018	^{1}0.0019	^{1}0.0020	^{1}0.0021	^{56}0.0007 $N^{0.067}$ ^{33}
60	MICROSOFT-6	^{4}0.0004	^{2}0.0004	^{2}0.0005	^{3}0.0006	^{3}0.0006	^{24}0.0001 $N^{0.185}$ ^{23}	^{1}0.0018	^{2}0.0019	^{2}0.0019	^{3}0.0021	^{4}0.0023	^{41}0.0005 $N^{0.091}$ ^{41}
61	NEC-0	^{80}0.0023	^{73}0.0030	^{42}0.0038	^{40}0.0047	^{37}0.0059	^{15}0.0000 $N^{0.324}$ ^{87}	^{79}0.0055	^{80}0.0064	^{64}0.0074	^{61}0.0085	^{55}0.0100	^{19}0.0003 $N^{0.306}$
62	NEC-1	^{119}0.0076	^{102}0.0080				^{101}0.0038 $N^{0.051}$ ^{12}	^{131}0.0135	^{132}0.0138	^{74}0.0142	^{70}0.0147	^{69}0.0154	^{107}0.0073 $N^{0.046}$ ^{16}
63	NEC-2	^{10}0.0008	^{8}0.0009	^{6}0.0009	^{5}0.0009	^{4}0.0009	^{75}0.0004 $N^{0.046}$ ^{21}	^{7}0.0022	^{5}0.0023	^{5}0.0023	^{5}0.0024	^{5}0.0025	^{72}0.0014 $N^{0.034}$ ^{13}
64	NEC-3	^{22}0.0011	^{21}0.0011	^{12}0.0011	^{9}0.0011	^{9}0.0011	^{87}0.0008 $N^{0.026}$ ^{2}	^{13}0.0026	^{12}0.0027	^{10}0.0028	^{7}0.0028	^{9}0.0029	^{81}0.0019 $N^{0.026}$ ^{3}
65	NEUROTECHNOLOGY-3	^{51}0.0038	^{51}0.0051				^{5}0.0000 $N^{0.326}$ ^{99}	^{95}0.0068	^{80}0.0083	^{88}0.0097	^{67}0.0116	^{67}0.0137	^{13}0.0008 $N^{0.245}$ ^{53}
66	NEUROTECHNOLOGY-4	^{31}0.0020	^{53}0.0024	^{30}0.0027	^{30}0.0031	^{27}0.0036	^{59}0.0002 $N^{0.187}$ ^{53}	^{60}0.0048	^{53}0.0051	^{45}0.0054	^{40}0.0057	^{36}0.0060	^{70}0.0016 $N^{0.081}$ ^{39}
67	NEUROTECHNOLOGY-5	^{40}0.0017	^{39}0.0018	^{31}0.0019	^{21}0.0021	^{16}0.0023	^{72}0.0004 $N^{0.108}$ ^{32}	^{51}0.0045	^{46}0.0047	^{35}0.0048	^{32}0.0050	^{29}0.0053	^{86}0.0021 $N^{0.065}$ ^{29}
68	NEWLAND-2							^{160}0.0235	^{144}0.0288	^{90}0.0332	^{78}0.0391		^{88}0.0011 $N^{0.177}$
69	NOBLIS-2	^{136}0.0366	^{131}0.0520				^{66}0.0002 $N^{0.353}$ ^{111}	^{161}0.0403	^{160}0.0560	^{88}0.0682	^{91}0.0940		^{20}0.0003 $N^{0.372}$ ^{110}
70	NTECHLAB-0	^{31}0.0013	^{32}0.0016	^{32}0.0021	^{23}0.0026	^{26}0.0032	^{13}0.0000 $N^{0.320}$ ^{85}	^{71}0.0033	^{30}0.0039	^{30}0.0043	0.0051	^{31}0.0058	^{22}0.0002 $N^{0.199}$ ^{63}
71	NTECHLAB-1	^{25}0.0013	^{33}0.0018	^{23}0.0022	^{18}0.0029	^{28}0.0038	^{9}0.0000 $N^{0.366}$	^{31}0.0034	^{35}0.0040				^{31}0.0003 $N^{0.177}$ ^{74}
72	NTECHLAB-3	^{21}0.0010	^{23}0.0012				^{22}0.0001 $N^{0.217}$ ^{59}	^{19}0.0028	^{19}0.0032	^{20}0.0035	^{22}0.0039	^{20}0.0044	^{30}0.0004 $N^{0.149}$ ^{53}

Table 14: **Investigation-mode: Effect of N on FNIR at rank 50** *For five enrollment population sizes, N, with T = 0 and FPIR = 1. The left five columns apply for consolidated enrollment of a variable number of lifetime images from each subject. The right five columns apply for enrollment of one recent image. Missing entries usually apply because another algorithm from the same developer was run instead. Some developers are missing because less accurate algorithms were not run on galleries with N > 1 600 000. Throughout blue superscripts indicate the rank of the algorithm for that column, and yellow highlighting indicates the most accurate value. Caution: The Power-low models are mostly intended to draw attention to the kind of behavior, not as a model to be used for prediction.*

MISSES NOT AT RANK 50 FNIR(N, T = 0, R = 50)		ENROL LIFETIME DATASET: FRVT 2018						ENROL MOST RECENT DATASET: FRVT 2018					
#	ALGORITHM	N=0.64M	N=1.6M	N=3.0M	N=6.0M	N=12.0M	aN^b	N=0.64M	N=1.6M	N=3.0M	N=6.0M	N=12.0M	aN^b
73	NTECHLAB-4	^{15}0.0009	^{15}0.0010	^{14}0.0012	^{13}0.0014	^{13}0.0016	^{26}0.0001 $N^{0.208\ 68}$	^{34}0.0027	^{33}0.0030	^{16}0.0032	^{16}0.0035	^{15}0.0039	^{43}0.0006 $N^{0.120\ 30}$
74	NTECHLAB-5	^{6}0.0007	^{7}0.0008				^{14}0.0000 $N^{0.237\ 66}$	^{6}0.0021	^{5}0.0025	^{8}0.0027	^{9}0.0031	^{11}0.0035	^{30}0.0002 $N^{0.166\ 39}$
75	NTECHLAB-6	^{5}0.0006	^{5}0.0008	^{5}0.0008	^{7}0.0010	^{9}0.0012	^{13}0.0000 $N^{0.244\ 68}$	^{5}0.0021	^{6}0.0023	^{7}0.0026	^{8}0.0028	^{7}0.0032	^{36}0.0008 $N^{0.147\ 32}$
76	QUANTASOFT-1	^{186}0.9843	^{149}0.9843				-	^{188}0.1140	^{184}0.1140	^{85}0.1140		^{77}0.1140	^{115}0.1140 $N^{0.000\ 1}$
77	RANKONE-0	^{112}0.0074	^{109}0.0100	^{55}0.0120	^{53}0.0146	^{34}0.0176	^{57}0.0001 $N^{0.237\ 80}$	^{130}0.0127	^{129}0.0159	^{77}0.0185	^{75}0.0206	^{73}0.0252	^{48}0.0006 $N^{0.226\ 59}$
78	RANKONE-1	^{98}0.0042	^{94}0.0055	^{51}0.0067	^{48}0.0082	^{42}0.0100	^{40}0.0001 $N^{0.301\ 92}$	^{104}0.0078	^{96}0.0086				^{83}0.0020 $N^{0.103\ 48}$
79	RANKONE-2	^{80}0.0037	^{85}0.0047				^{53}0.0001 $N^{0.258\ 72}$	^{98}0.0075	^{100}0.0087	^{87}0.0098	^{66}0.0111	^{66}0.0128	^{51}0.0006 $N^{0.184\ 77}$
80	RANKONE-3	^{81}0.0037	^{83}0.0047	^{46}0.0055	^{44}0.0067	^{41}0.0079	^{59}0.0001 $N^{0.258\ 75}$	^{97}0.0075	^{99}0.0087	^{86}0.0098	^{65}0.0111	^{65}0.0128	^{90}0.0006 $N^{0.184\ 76}$
81	RANKONE-4	^{107}0.0058	^{100}0.0079				^{35}0.0001 $N^{0.335\ 91}$	^{116}0.0099	^{120}0.0128	^{78}0.0153			^{19}0.0002 $N^{0.264\ 100}$
82	RANKONE-5	^{86}0.0021	^{55}0.0025	^{36}0.0029	^{35}0.0034	^{29}0.0040	^{49}0.0001 $N^{0.220\ 80}$	^{74}0.0053	^{75}0.0058	^{55}0.0063	^{53}0.0069	^{48}0.0077	^{62}0.0009 $N^{0.127\ 56}$
83	REALNETWORKS-0	^{106}0.0059	^{104}0.0083	^{54}0.0108			^{16}0.0000 $N^{0.393\ 102}$	^{105}0.0077	^{110}0.0098				^{17}0.0002 $N^{0.267\ 98}$
84	REALNETWORKS-2	^{96}0.0042	^{96}0.0061				^{10}0.0000 $N^{0.423\ 104}$	^{99}0.0075	^{108}0.0098	^{72}0.0119	^{71}0.0149	^{70}0.0155	^{21}0.0002 $N^{0.362\ 97}$
85	REMARKAI-2	^{34}0.0013	^{33}0.0016				^{37}0.0001 $N^{0.234\ 81}$	^{38}0.0038	^{37}0.0042	^{33}0.0046	^{28}0.0050		^{57}0.0007 $N^{0.125\ 53}$
86	SENSETIME-0	^{29}0.0012	^{26}0.0013				-	^{41}0.0041	^{38}0.0041	^{28}0.0042	^{25}0.0043	^{19}0.0044	^{95}0.0028 $N^{0.026\ 4}$
87	SENSETIME-1	^{28}0.0011	^{22}0.0012				-	^{40}0.0040	^{37}0.0041	^{27}0.0041	^{24}0.0042	^{26}0.0048	^{80}0.0018 $N^{0.057\ 15}$
88	SHAMAN-3	^{152}0.0344	^{127}0.0404	^{62}0.0452			^{38}0.0032 $N^{0.177\ 49}$	^{167}0.0468	^{161}0.0544				^{105}0.0053 $N^{0.163\ 67}$
89	SHAMAN-7	^{147}0.0243	^{120}0.0248				^{107}0.0183 $N^{0.021\ 8}$	^{160}0.0334	^{149}0.0339	^{81}0.0344	^{77}0.0352	^{74}0.0362	^{113}0.0230 $N^{0.028\ 7}$
90	SIAT-1	^{183}0.2635	^{146}0.2635	^{68}0.2636			^{110}0.2626 $N^{0.000\ 2}$	^{22}0.0029	^{14}0.0030	^{13}0.0031	^{11}0.0032	^{9}0.0033	^{74}0.0016 $N^{0.046\ 17}$
91	SIAT-2	^{181}0.2124	^{142}0.2124				^{109}0.2116 $N^{0.000\ 3}$	^{26}0.0031	^{22}0.0032	^{17}0.0032	^{13}0.0033	^{10}0.0034	^{84}0.0020 $N^{0.032\ 11}$
92	SMILART-4	^{188}0.8160	^{148}0.9522				^{108}0.0859 $N^{0.168\ 46}$	^{200}0.9159	^{193}0.9638	^{88}0.9906			^{116}0.4632 $N^{0.051\ 19}$
93	SYNESIS-3	^{166}0.0582	^{132}0.0632				^{106}0.0174 $N^{0.090\ 28}$	^{175}0.0851	^{171}0.0891	^{84}0.0942	^{80}0.1020	^{76}0.1126	^{114}0.0231 $N^{0.095\ 32}$
94	TEVIAN-4	^{47}0.0019	^{45}0.0022	^{27}0.0025			^{58}0.0002 $N^{0.185\ 51}$	^{42}0.0041	^{44}0.0046				^{97}0.0006 $N^{0.142\ 61}$
95	TEVIAN-5	^{36}0.0014	^{34}0.0017				^{82}0.0002 $N^{0.160\ 44}$	^{30}0.0034	^{32}0.0037	^{26}0.0041	^{27}0.0044	^{28}0.0050	^{94}0.0006 $N^{0.139\ 57}$
96	TIGER-0	^{109}0.0061	^{108}0.0097	^{56}0.0125	^{54}0.0164		^{11}0.0000 $N^{0.464\ 105}$	^{177}0.0098	^{123}0.0139				^{7}0.0001 $N^{0.384\ 112}$
97	TIGER-2	^{18}0.0010	^{21}0.0012				^{27}0.0001 $N^{0.208\ 67}$	^{37}0.0028	^{38}0.0030	^{19}0.0034	^{17}0.0038	^{22}0.0045	^{29}0.0003 $N^{0.161\ 56}$
98	TONGYITRANS-1	^{105}0.0057	^{95}0.0060	^{49}0.0062	^{45}0.0067		^{97}0.0020 $N^{0.076\ 22}$	^{65}0.0049	^{59}0.0052				^{88}0.0022 $N^{0.061\ 27}$
99	TOSHIBA-0	^{29}0.0011	^{24}0.0012				^{65}0.0002 $N^{0.126\ 37}$	^{54}0.0037	^{33}0.0039	^{36}0.0041	^{25}0.0048	^{64}0.0127	^{5}0.0000 $N^{0.360\ 107}$
100	VD-0	^{173}0.1006	^{139}0.1421	^{66}0.1752	^{62}0.2147		^{87}0.0011 $N^{0.340\ 92}$	^{180}0.1248	^{185}0.1699				^{71}0.0014 $N^{0.336\ 105}$
101	VD-1	^{126}0.0098	^{110}0.0105				^{98}0.0031 $N^{0.085\ 24}$	^{134}0.0145	^{128}0.0155	^{76}0.0166	^{74}0.0179	^{72}0.0196	^{99}0.0036 $N^{0.103\ 47}$
102	VIGILANTSOLUTIONS-3	^{114}0.0072	^{111}0.0110	^{57}0.0143	^{52}0.0143		^{46}0.0001 $N^{0.322\ 86}$	^{127}0.0118	^{131}0.0166				^{3}0.0001 $N^{0.373\ 111}$
103	VISIONLABS-3	^{80}0.0030	^{82}0.0042	^{50}0.0066	^{51}0.0119		^{5}0.0000 $N^{0.612\ 107}$	^{87}0.0057	^{90}0.0073	^{70}0.0106	^{73}0.0166		^{4}0.0000 $N^{0.481\ 113}$
104	VISIONLABS-4	^{20}0.0010	^{19}0.0011				^{68}0.0002 $N^{0.103\ 31}$	^{11}0.0025	^{11}0.0027	^{12}0.0030	^{15}0.0039	^{34}0.0059	^{6}0.0000 $N^{0.280\ 103}$
105	VISIONLABS-5	^{17}0.0009	^{14}0.0010	^{15}0.0012	^{16}0.0016	^{18}0.0026	^{8}0.0000 $N^{0.341\ 94}$	^{10}0.0025	^{10}0.0026	^{11}0.0029	^{14}0.0033	^{18}0.0044	^{13}0.0002 $N^{0.192\ 80}$
106	VISIONLABS-6	^{19}0.0010	^{16}0.0010				^{77}0.0005 $N^{0.056\ 15}$	^{9}0.0023	^{8}0.0025	^{9}0.0027	^{10}0.0031	^{16}0.0040	^{16}0.0002 $N^{0.177\ 75}$
107	VISIONLABS-7	^{16}0.0009	^{13}0.0010	^{9}0.0010	^{8}0.0011	^{7}0.0011	^{70}0.0004 $N^{0.070\ 19}$	^{8}0.0023	^{7}0.0024	^{6}0.0025	^{5}0.0025	^{8}0.0032	^{46}0.0006 $N^{0.195\ 84}$
108	VOCORD-3	^{69}0.0023	^{59}0.0025	^{38}0.0028	^{28}0.0031		^{76}0.0004 $N^{0.133\ 36}$	^{39}0.0040	^{41}0.0042				^{78}0.0017 $N^{0.069\ 29}$
109	VOCORD-5	^{76}0.0027	^{69}0.0029				^{92}0.0013 $N^{0.056\ 14}$	^{71}0.0051	^{66}0.0054	^{45}0.0056	^{40}0.0060	^{42}0.0064	^{82}0.0019 $N^{0.074\ 35}$
110	YISHENG-1	^{84}0.0035	^{88}0.0047	^{47}0.0058	^{46}0.0072		^{19}0.0000 $N^{0.315\ 88}$	^{96}0.0069	^{95}0.0092				^{42}0.0005 $N^{0.191\ 79}$
111	YITU-0	^{72}0.0026	^{63}0.0027	^{37}0.0029	^{32}0.0031	^{26}0.0034	^{84}0.0008 $N^{0.090\ 26}$	^{63}0.0048	^{52}0.0049	^{38}0.0052	^{38}0.0054	^{30}0.0057	^{87}0.0021 $N^{0.080\ 36}$
112	YITU-1	^{70}0.0026	^{61}0.0027	^{34}0.0029	^{29}0.0031	^{25}0.0034	^{83}0.0008 $N^{0.090\ 27}$	^{62}0.0048	^{51}0.0049				^{97}0.0033 $N^{0.029\ 9}$
113	YITU-2	^{12}0.0008	^{9}0.0009	^{7}0.0009	^{6}0.0010	^{5}0.0010	^{71}0.0004 $N^{0.068\ 18}$	^{32}0.0034	^{29}0.0035	^{34}0.0036	^{18}0.0036	^{13}0.0037	^{89}0.0024 $N^{0.027\ 6}$
114	YITU-3	^{45}0.0018	^{37}0.0018				^{89}0.0011 $N^{0.036\ 10}$	^{54}0.0046	^{46}0.0047	^{34}0.0047	^{31}0.0048	^{27}0.0049	^{96}0.0031 $N^{0.029\ 8}$
115	YITU-4	^{7}0.0008	^{9}0.0008	^{4}0.0008	^{4}0.0008	^{6}0.0011	^{80}0.0006 $N^{0.015\ 4}$	^{28}0.0032	^{24}0.0033	^{18}0.0033	^{12}0.0033	^{35}0.0060	^{88}0.0008 $N^{0.162\ 70}$
116	YITU-5	^{41}0.0017	^{39}0.0017	^{16}0.0017	^{18}0.0017	^{17}0.0018	^{78}0.0014 $N^{0.015\ 5}$	^{47}0.0044	^{42}0.0044	^{31}0.0044	^{37}0.0044	^{31}0.0045	^{108}0.0039 $N^{0.018\ 2}$

Table 15: **Investigation-mode: Effect of N on FNIR at rank 50** *For five enrollment population sizes, N, with T = 0 and FPIR = 1. The left five columns apply for consolidated enrollment of a variable number of lifetime images from each subject. The right five columns apply for enrollment of one recent image. Missing entries usually apply because another algorithm from the same developer was run instead. Some developers are missing because less accurate algorithms were not run on galleries with N > 1 600 000. Throughout blue superscripts indicate the rank of the algorithm for that column, and yellow highlighting indicates the most accurate value. Caution: The Power-low models are mostly intended to draw attention to the kind of behavior, not as a model to be used for prediction.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

MISSES OUTSIDE RANK R FNIR(N, T=0, R)		RESOURCE USAGE TEMPLATE		ENROLL LIFETIME CONSOLIDATED = 1.6M FRVT 2018 MUGSHOTS				ENROL MOST RECENT, N = 1.6M			
#	ALGORITHM	BYTES	MSEC	R=1	R=10	R=50	WORK-10	R=1	R=10	R=50	WORK-10
1	3DIVI-0	^{184}4096	^{90}426				^{199}10.000	^{118}0.0344	^{118}0.0344	^{118}0.0344	^{117}1.190
2	3DIVI-1	^{195}4224	^{94}428				^{198}10.000	^{119}0.0375	^{119}0.0375	^{119}0.0375	^{123}1.233
3	3DIVI-2	^{49}528	^{92}428				^{177}10.000	^{124}0.0404	^{124}0.0404	^{124}0.0404	^{129}1.259
4	3DIVI-3	^{42}512	^{130}625	^{123}0.0645	^{123}0.0645	^{123}0.0645	^{121}1.345	^{152}0.0857	^{152}0.0857	^{152}0.0857	^{148}1.469
5	3DIVI-4	^{177}4096	^{131}628	^{85}0.0133	^{85}0.0133	^{85}0.0133	^{82}1.069	^{96}0.0201	^{96}0.0201	^{96}0.0201	^{94}1.115
6	3DIVI-5	^{182}4096	^{132}653	^{86}0.0133	^{86}0.0133	^{86}0.0133	^{81}1.069	^{97}0.0202	^{97}0.0202	^{97}0.0202	^{95}1.116
7	3DIVI-6	^{48}528	^{136}653	^{95}0.0186	^{95}0.0186	^{95}0.0186	^{101}1.127	^{110}0.0265	^{110}0.0265	^{110}0.0265	^{115}1.186
8	ALCHERA-0	^{111}2048	^{42}263	^{82}0.0121	^{82}0.0121	^{82}0.0121	^{87}1.085	^{92}0.0186	^{92}0.0186	^{92}0.0186	^{104}1.138
9	ALCHERA-1	^{120}2048	^{8}66	^{148}0.9824	^{148}0.9824	^{148}0.9824	^{148}9.748	^{198}0.9869	^{198}0.9869	^{198}0.9869	^{199}9.812
10	ALCHERA-2	^{131}2048	^{18}115	^{124}0.0914	^{124}0.0914	^{124}0.0914	^{125}1.552	^{153}0.0973	^{153}0.0973	^{153}0.0973	^{152}1.567
11	ALCHERA-3	^{132}2048	^{137}548	^{91}0.0159	^{91}0.0159	^{91}0.0159	^{88}1.066	^{72}0.0127	^{72}0.0127	^{72}0.0127	^{66}1.074
12	ANKE-0	^{165}2072	^{96}431	^{72}0.0100	^{72}0.0100	^{72}0.0100	^{68}1.055	^{86}0.0158	^{86}0.0158	^{86}0.0158	^{80}1.095
13	ANKE-1	^{164}2072	^{97}433	^{73}0.0101	^{73}0.0101	^{73}0.0101	^{67}1.055	^{87}0.0158	^{87}0.0158	^{87}0.0158	^{82}1.096
14	AWARE-0	^{100}1564	^{140}653				^{192}10.000	^{145}0.0639	^{145}0.0639	^{145}0.0639	^{147}1.439
15	AWARE-1	^{99}1564	^{136}651				^{193}10.000	^{141}0.0587	^{141}0.0587	^{141}0.0587	^{143}1.382
16	AWARE-2	^{167}2076	^{197}912				^{194}10.000	^{142}0.0600	^{142}0.0600	^{142}0.0600	^{145}1.416
17	AWARE-3	^{166}2076	^{163}716	^{101}0.0209	^{101}0.0209	^{101}0.0209	^{99}1.110	^{116}0.0332	^{116}0.0332	^{116}0.0332	^{116}1.186
18	AWARE-4	^{2}92	^{160}712	^{119}0.0529	^{119}0.0529	^{119}0.0529	^{115}1.275	^{147}0.0704	^{147}0.0704	^{147}0.0704	^{141}1.378
19	AWARE-5	^{175}3100	^{180}817	^{100}0.0208	^{100}0.0208	^{100}0.0208	^{98}1.110	^{117}0.0337	^{117}0.0337	^{117}0.0337	^{118}1.191
20	AWARE-6	^{3}124	^{175}818	^{120}0.0538	^{120}0.0538	^{120}0.0538	^{117}1.286	^{149}0.0722	^{149}0.0722	^{149}0.0722	^{144}1.394
21	AYONIX-0	^{81}1036	^{1}10	^{144}0.4649	^{144}0.4649	^{144}0.4649	^{144}4.268	^{191}0.4519	^{191}0.4519	^{191}0.4519	^{192}4.304
22	AYONIX-1	^{82}1036	^{3}12	^{140}0.3364	^{140}0.3364	^{140}0.3364	^{139}3.073	^{187}0.3432	^{187}0.3432	^{187}0.3432	^{186}3.244
23	AYONIX-2	^{77}1036	^{2}11	^{137}0.2606	^{137}0.2606	^{137}0.2606	^{136}2.620	^{186}0.3432	^{186}0.3432	^{186}0.3432	^{187}3.244
24	CAMVI-1	^{62}1024	^{24}177				^{156}10.000	^{179}0.2267	^{179}0.2267	^{179}0.2267	^{176}2.419
25	CAMVI-2	^{71}1024	^{172}774				^{179}10.000	^{160}0.1292	^{160}0.1292	^{160}0.1292	^{159}1.781
26	CAMVI-3	^{73}1024	^{158}707	^{112}0.0368	^{112}0.0368	^{112}0.0368	^{119}1.330	^{140}0.0544	^{140}0.0544	^{140}0.0544	^{150}1.488
27	CAMVI-4	^{60}1024	^{166}718	^{110}0.0326	^{110}0.0326	^{110}0.0326	^{118}1.291	^{137}0.0490	^{137}0.0490	^{137}0.0490	^{146}1.438
28	CAMVI-5	^{66}1024	^{170}769	^{116}0.0458	^{116}0.0458	^{116}0.0458	^{123}1.410	^{146}0.0673	^{146}0.0673	^{146}0.0673	^{153}1.602
29	COGENT-0	^{44}525	^{118}551	^{77}0.0106	^{77}0.0106	^{77}0.0106	^{76}1.062	^{74}0.0131	^{74}0.0131	^{74}0.0131	^{81}1.111
30	COGENT-1	^{45}525	^{119}552	^{76}0.0106	^{76}0.0106	^{76}0.0106	^{77}1.062	^{73}0.0131	^{73}0.0131	^{73}0.0131	^{80}1.111
31	COGENT-2	^{84}1043	^{208}987	^{30}0.0027	^{30}0.0027	^{30}0.0027	^{22}1.017	^{26}0.0062	^{26}0.0062	^{26}0.0062	^{30}1.045
32	COGENT-3	^{85}1043	^{202}960	^{29}0.0037	^{29}0.0037	^{29}0.0037	^{32}1.024	^{27}0.0064	^{27}0.0064	^{27}0.0064	^{35}1.047
33	COGNITEC-0	^{154}2052	^{25}176	^{96}0.0189	^{96}0.0189	^{96}0.0189	^{93}1.108	^{112}0.0278	^{112}0.0278	^{112}0.0278	^{111}1.160
34	COGNITEC-1	^{150}2052	^{28}202	^{66}0.0089	^{66}0.0089	^{66}0.0089	^{64}1.048	^{83}0.0143	^{83}0.0143	^{83}0.0143	^{76}1.086
35	COGNITEC-2	^{157}2052	^{34}227	^{34}0.0044	^{34}0.0044	^{34}0.0044	^{35}1.027	^{42}0.0083	^{42}0.0083	^{42}0.0083	^{46}1.059
36	COGNITEC-3	^{148}2052	^{50}297	^{37}0.0048	^{37}0.0048	^{37}0.0048	^{40}1.031	^{45}0.0088	^{45}0.0088	^{45}0.0088	^{52}1.062
37	DAHUA-0	^{144}2048	^{72}378	^{56}0.0070	^{56}0.0070	^{56}0.0070	^{62}1.047	^{64}0.0115	^{64}0.0115	^{64}0.0115	^{73}1.082
38	DAHUA-1	^{134}2048	^{68}371	^{40}0.0049	^{40}0.0049	^{40}0.0049	^{38}1.030	^{47}0.0089	^{47}0.0089	^{47}0.0089	^{48}1.058
39	DERMALOG-0	^{5}128	^{64}344				^{170}10.000	^{161}0.1309	^{161}0.1309	^{161}0.1309	^{158}1.779
40	DERMALOG-1	^{7}128	^{22}171				^{200}10.000	^{163}0.1563	^{163}0.1563	^{163}0.1563	^{163}1.945
41	DERMALOG-2	^{23}256	^{65}344				^{175}10.000	^{162}0.1377	^{162}0.1377	^{162}0.1377	^{161}1.817
42	DERMALOG-3	^{8}128	^{31}211	^{128}0.0970	^{128}0.0970	^{128}0.0970	^{127}1.566	^{158}0.1281	^{158}0.1281	^{158}0.1281	^{157}1.752
43	DERMALOG-4	^{4}128	^{28}208	^{126}0.0961	^{126}0.0961	^{126}0.0961	^{126}1.561	^{157}0.1274	^{157}0.1274	^{157}0.1274	^{156}1.748
44	DERMALOG-5	^{6}128	^{109}532	^{79}0.0113	^{79}0.0113	^{79}0.0113	^{91}1.089	^{89}0.0171	^{89}0.0171	^{89}0.0171	^{103}1.137
45	DERMALOG-6	^{15}256	^{105}514	^{48}0.0060	^{48}0.0060	^{48}0.0060	^{63}1.047	^{56}0.0102	^{56}0.0102	^{56}0.0102	^{72}1.081
46	EVERAI-0	^{124}2048	^{98}438	^{93}0.0166	^{93}0.0166	^{93}0.0166	^{103}1.141	^{99}0.0209	^{99}0.0209	^{99}0.0209	^{112}1.174
47	EVERAI-1	^{116}2048	^{125}590	^{21}0.0027	^{21}0.0027	^{21}0.0027	^{21}1.017	^{20}0.0056	^{20}0.0056	^{20}0.0056	^{19}1.038
48	EVERAI-2	^{139}2048	^{71}377	^{22}0.0029	^{22}0.0029	^{22}0.0029	^{26}1.018	^{22}0.0058	^{22}0.0058	^{22}0.0058	^{21}1.039
49	EVERAI-3	^{110}2048	^{167}735	^{16}0.0023	^{16}0.0023	^{16}0.0023	^{17}1.016	^{15}0.0047	^{15}0.0047	^{15}0.0047	^{14}1.034
50	EYEDEA-0	^{190}4152	^{89}424				^{184}10.000	^{184}0.3000	^{184}0.3000	^{184}0.3000	^{184}2.864
51	EYEDEA-1	^{80}1036	^{50}311				^{185}10.000	^{172}0.1981	^{172}0.1981	^{172}0.1981	^{171}2.226
52	EYEDEA-2	^{78}1036	^{95}429				^{187}10.000	^{173}0.2000	^{173}0.2000	^{173}0.2000	^{172}2.246
53	EYEDEA-3	^{79}1036	^{73}385	^{122}0.0613	^{122}0.0613	^{122}0.0613	^{120}1.348	^{151}0.0824	^{151}0.0824	^{151}0.0824	^{149}1.470
54	GLORY-0	^{33}418	^{16}160	^{130}0.1336	^{130}0.1336	^{130}0.1336	^{131}1.965	^{168}0.1803	^{168}0.1803	^{168}0.1803	^{173}2.318
55	GLORY-1	^{112}1726	^{91}405	^{125}0.0932	^{125}0.0932	^{125}0.0932	^{129}1.656	^{159}0.1291	^{159}0.1291	^{159}0.1291	^{162}1.925
56	GORILLA-0	^{206}8300	^{91}427				^{178}10.000				^{200}10.000
57	GORILLA-1	^{170}2156	^{21}169	^{114}0.0414	^{114}0.0414	^{114}0.0414	^{110}1.211	^{143}0.0627	^{143}0.0627	^{143}0.0627	^{136}1.351
58	GORILLA-2	^{87}1132	^{62}341	^{87}0.0137	^{87}0.0137	^{87}0.0137	^{80}1.067	^{100}0.0220	^{100}0.0220	^{100}0.0220	^{96}1.116
59	GORILLA-3	^{169}2156	^{129}563	^{103}0.0245	^{103}0.0245	^{103}0.0245	^{97}1.110	^{121}0.0384	^{121}0.0384	^{121}0.0384	^{114}1.178
60	HBINNO-0	^{46}520	^{40}265				^{181}10.000	^{183}0.2746	^{183}0.2746	^{183}0.2746	^{183}2.743
61	HIK-0	^{115}1808	^{194}875				^{151}10.000	^{107}0.0236	^{107}0.0236	^{107}0.0236	^{113}1.176
62	HIK-1	^{117}1808	^{176}820				^{186}10.000	^{91}0.0173	^{91}0.0173	^{91}0.0173	^{97}1.116
63	HIK-2	^{106}1808	^{178}820	^{94}0.0185	^{94}0.0185	^{94}0.0185	^{100}1.119	^{90}0.0172	^{90}0.0172	^{90}0.0172	^{93}1.115
64	HIK-3	^{90}1408	^{132}633	^{78}0.0107	^{78}0.0107	^{78}0.0107	^{74}1.057	^{82}0.0141	^{82}0.0141	^{82}0.0141	^{74}1.082
65	HIK-4	^{89}1152	^{104}510	^{75}0.0104	^{75}0.0104	^{75}0.0104	^{70}1.055	^{80}0.0138	^{80}0.0138	^{80}0.0138	^{70}1.081
66	HIK-5	^{89}1408	^{127}619	^{25}0.0034	^{25}0.0034	^{25}0.0034	^{25}1.018	^{29}0.0067	^{29}0.0067	^{29}0.0067	^{27}1.043
67	HIK-6	^{91}1408	^{126}610	^{27}0.0034	^{27}0.0034	^{27}0.0034	^{24}1.018	^{30}0.0067	^{30}0.0067	^{30}0.0067	^{28}1.043
68	IDEMIA-0	^{31}364	^{89}416	^{52}0.0063	^{52}0.0063	^{52}0.0063	^{47}1.034	^{61}0.0113	^{61}0.0113	^{61}0.0113	^{61}1.070
69	IDEMIA-1	^{32}364	^{88}417	^{53}0.0065	^{53}0.0065	^{53}0.0065	^{49}1.035	^{65}0.0116	^{65}0.0116	^{65}0.0116	^{63}1.072
70	IDEMIA-2	^{30}364	^{85}417	^{70}0.0099	^{70}0.0099	^{70}0.0099	^{71}1.056	^{71}0.0126	^{71}0.0126	^{71}0.0126	^{71}1.081
71	IDEMIA-3	^{50}528	^{149}689	^{45}0.0054	^{45}0.0054	^{45}0.0054	^{37}1.033	^{54}0.0095	^{54}0.0095	^{54}0.0095	^{57}1.066
72	IDEMIA-4	^{47}528	^{147}669	^{43}0.0052	^{43}0.0052	^{43}0.0052	^{33}1.029	^{50}0.0092	^{50}0.0092	^{50}0.0092	^{49}1.061

Table 16: **Rank-based accuracy for the FRVT 2018 mugshot sets.** *In columns 3 and 4 are template size and template generation duration. Thereafter values are rank-based FNIR with T = 0 and FPIR = 1. This is appropriate to investigational uses but not those with higher volumes where candidates from all searches would need review. Columns 5 - 9 show FRVT 2018 accuracy for various ranks for galleries unenrolled with all lifetime images. Column 10 is a workload statistic, a small value shows an algorithm front-loads mates into the first 10 candidates. The last four columns gives analogous results for enrollment only of the most recent image - see Figure 8. Throughout, blue superscripts indicate the rank of the algorithm for that column, and the best value is highlighted in yellow.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

MISSES OUTSIDE RANK R		RESOURCE USAGE		ENROLL LIFETIME CONSOLIDATED = 1.6M				ENROL MOST RECENT, N = 1.6M			
FNIR(N, T=0, R)		TEMPLATE		FRVT 2018 MUGSHOTS							
#	ALGORITHM	BYTES	MSEC	R=1	R=10	R=50	WORK-10	R=1	R=10	R=50	WORK-10
73	IDEMIA-5	^{29}352	^{70}374	^{80}0.0062	^{80}0.0062	^{80}0.0062	^{46}1.034	^{99}0.0107	^{99}0.0107	^{99}0.0107	^{88}1.068
74	IDEMIA-6	^{28}352	^{69}373	^{97}0.0071	^{97}0.0071	^{97}0.0071	^{53}1.039	^{89}0.0122	^{89}0.0122	^{89}0.0122	^{87}1.076
75	IMAGUS-0	^{43}512	^{3}43				^{198}10.000	^{185}0.3054	^{185}0.3054	^{185}0.3054	^{188}2.977
76	IMAGUS-2	^{37}512	^{9}76	^{133}0.1833	^{133}0.1833	^{133}0.1833	^{182}2.070	^{177}0.2223	^{177}0.2223	^{177}0.2223	^{174}2.329
77	IMAGUS-3	^{41}512	^{7}57	^{139}0.3008	^{139}0.3008	^{139}0.3008	^{138}2.951	^{188}0.3576	^{188}0.3576	^{188}0.3576	^{188}3.380
78	INCODE-0	^{97}1024	^{26}190	^{113}0.0376	^{113}0.0376	^{113}0.0376	^{109}1.201	^{139}0.0515	^{139}0.0515	^{139}0.0515	^{132}1.285
79	INCODE-1	^{127}2048	^{151}690	^{84}0.0131	^{84}0.0131	^{84}0.0131	^{78}1.066	^{93}0.0190	^{93}0.0190	^{93}0.0190	^{88}1.106
80	INCODE-2	^{122}2048	^{49}291	^{81}0.0120	^{81}0.0120	^{81}0.0120	^{75}1.060	^{98}0.0203	^{98}0.0203	^{98}0.0203	^{92}1.113
81	INCODE-3	^{117}2048	^{155}704	^{65}0.0088	^{65}0.0088	^{65}0.0088	^{60}1.044	^{85}0.0153	^{85}0.0153	^{85}0.0153	^{75}1.086
82	INNOVATRICS-0	^{54}530	^{100}455				^{186}10.000	^{127}0.0421	^{127}0.0421	^{127}0.0421	^{135}1.234
83	INNOVATRICS-1	^{52}530	^{58}316				^{182}10.000	^{126}0.0421	^{126}0.0421	^{126}0.0421	^{134}1.234
84	INNOVATRICS-2	^{53}530	^{41}255	^{118}0.0499	^{118}0.0499	^{118}0.0499	^{122}1.354	^{136}0.0475	^{136}0.0475	^{136}0.0475	^{140}1.343
85	INNOVATRICS-3	^{51}530	^{40}255	^{104}0.0301	^{104}0.0301	^{104}0.0301	^{104}1.147	^{113}0.0287	^{113}0.0287	^{113}0.0287	^{118}1.151
86	INNOVATRICS-4	^{95}1076	^{80}406	^{61}0.0081	^{61}0.0081	^{61}0.0081	^{59}1.042	^{84}0.0149	^{84}0.0149	^{84}0.0149	^{77}1.087
87	ISYSTEMS-0	^{141}2048	^{33}222	^{64}0.0085	^{64}0.0085	^{64}0.0085	^{74}1.059	^{77}0.0136	^{77}0.0136	^{77}0.0136	^{83}1.098
88	ISYSTEMS-1	^{84}1024	^{32}222	^{63}0.0085	^{63}0.0085	^{63}0.0085	^{73}1.058	^{76}0.0136	^{76}0.0136	^{76}0.0136	^{84}1.098
89	ISYSTEMS-2	^{139}2048	^{59}316	^{37}0.0046	^{37}0.0046	^{37}0.0046	^{44}1.032	^{44}0.0088	^{44}0.0088	^{44}0.0088	^{53}1.062
90	ISYSTEMS-3	^{119}2048	^{159}856	^{32}0.0040	^{32}0.0040	^{32}0.0040	^{37}1.029	^{37}0.0075	^{37}0.0075	^{37}0.0075	^{49}1.057
91	LOOKMAN-3	^{26}292	^{63}342	^{67}0.0089	^{67}0.0089	^{67}0.0089	^{65}1.074	^{62}0.0114	^{62}0.0114	^{62}0.0114	^{79}1.095
92	LOOKMAN-4	^{55}548	^{60}325	^{68}0.0091	^{68}0.0091	^{68}0.0091	^{66}1.074	^{66}0.0117	^{66}0.0117	^{66}0.0117	^{81}1.096
93	MEGVII-0	^{118}2048	^{174}794	^{71}0.0099	^{71}0.0099	^{71}0.0099	^{65}1.048	^{51}0.0094	^{51}0.0094	^{51}0.0094	^{39}1.052
94	MEGVII-1	^{178}4096	^{137}652				^{189}10.000	^{78}0.0137	^{78}0.0137	^{78}0.0137	^{86}1.102
95	MEGVII-2	^{180}4096	^{140}656				^{176}10.000	^{79}0.0137	^{79}0.0137	^{79}0.0137	^{87}1.102
96	MICROFOCUS-0	^{15}256	^{106}526				^{168}10.000	^{195}0.5972	^{195}0.5972	^{195}0.5972	^{195}5.397
97	MICROFOCUS-1	^{24}256	^{107}527				^{180}10.000	^{196}0.5972	^{196}0.5972	^{196}0.5972	^{196}5.398
98	MICROFOCUS-2	^{22}256	^{108}529				^{174}10.000	^{197}0.6272	^{197}0.6272	^{197}0.6272	^{197}5.839
99	MICROFOCUS-3	^{21}256	^{46}269	^{146}0.5389	^{146}0.5389	^{146}0.5389	^{146}4.849	^{194}0.5953	^{194}0.5953	^{194}0.5953	^{194}5.373
100	MICROFOCUS-4	^{25}256	^{47}270	^{145}0.5191	^{145}0.5191	^{145}0.5191	^{145}4.688	^{193}0.5775	^{193}0.5775	^{193}0.5775	^{193}5.212
101	MICROFOCUS-5	^{16}256	^{45}266	^{141}0.3701	^{141}0.3701	^{141}0.3701	^{141}3.437	^{189}0.4257	^{189}0.4257	^{189}0.4257	^{189}3.877
102	MICROFOCUS-6	^{17}256	^{44}265	^{142}0.3732	^{142}0.3732	^{142}0.3732	^{142}3.453	^{190}0.4283	^{190}0.4283	^{190}0.4283	^{190}3.897
103	MICROSOFT-0	^{36}512	^{38}283	^{19}0.0026	^{19}0.0026	^{19}0.0026	^{16}1.015	^{23}0.0058	^{23}0.0058	^{23}0.0058	^{20}1.038
104	MICROSOFT-1	^{74}1024	^{66}349	^{18}0.0026	^{18}0.0026	^{18}0.0026	^{15}1.015	^{21}0.0056	^{21}0.0056	^{21}0.0056	^{18}1.038
105	MICROSOFT-2	^{63}1024	^{120}565	^{23}0.0029	^{23}0.0029	^{23}0.0029	^{20}1.016	^{25}0.0061	^{25}0.0061	^{25}0.0061	^{25}1.041
106	MICROSOFT-3	^{76}1024	^{80}404	^{4}0.0011	^{4}0.0011	^{4}0.0011	^{2}1.007	^{4}0.0032	^{4}0.0032	^{4}0.0032	^{3}1.022
107	MICROSOFT-4	^{123}2048	^{171}773	^{1}0.0010	^{1}0.0010	^{1}0.0010	^{1}1.006	^{2}0.0031	^{2}0.0031	^{2}0.0031	^{2}1.022
108	MICROSOFT-5	^{70}1024	^{148}673	^{5}0.0013	^{5}0.0013	^{5}0.0013	^{3}1.007	^{5}0.0033	^{5}0.0033	^{5}0.0033	^{1}1.021
109	MICROSOFT-6	^{68}1024	^{152}695	^{7}0.0014	^{7}0.0014	^{7}0.0014	^{4}1.007	^{6}0.0033	^{6}0.0033	^{6}0.0033	^{4}1.023
110	NEC-0	^{172}2592	^{10}82	^{83}0.0127	^{83}0.0127	^{83}0.0127	^{79}1.066	^{94}0.0196	^{94}0.0196	^{94}0.0196	^{89}1.110
111	NEC-1	^{171}2592	^{11}88	^{92}0.0164	^{92}0.0164	^{92}0.0164	^{92}1.101	^{106}0.0235	^{106}0.0235	^{106}0.0235	^{110}1.158
112	NEC-2	^{101}1616	^{141}653	^{3}0.0011	^{3}0.0011	^{3}0.0011	^{6}1.009	^{1}0.0028	^{1}0.0028	^{1}0.0028	^{5}1.023
113	NEC-3	^{102}1712	^{160}690	^{6}0.0013	^{6}0.0013	^{6}0.0013	^{9}1.011	^{3}0.0031	^{3}0.0031	^{3}0.0031	^{8}1.026
114	NEUROTECHNOLOGY-0	^{197}5214	^{154}702				^{171}10.000	^{138}0.0497	^{138}0.0497	^{138}0.0497	^{131}1.278
115	NEUROTECHNOLOGY-1	^{199}5214	^{146}661				^{196}10.000	^{135}0.0467	^{135}0.0467	^{135}0.0467	^{128}1.250
116	NEUROTECHNOLOGY-2	^{198}5214	^{144}658				^{193}10.000	^{134}0.0465	^{134}0.0465	^{134}0.0465	^{127}1.249
117	NEUROTECHNOLOGY-3	^{136}2048	^{116}547	^{98}0.0199	^{98}0.0199	^{98}0.0199	^{95}1.108	^{109}0.0250	^{109}0.0250	^{109}0.0250	^{106}1.148
118	NEUROTECHNOLOGY-4	^{142}2048	^{115}543	^{47}0.0058	^{47}0.0058	^{47}0.0058	^{57}1.037	^{40}0.0082	^{40}0.0082	^{40}0.0082	^{34}1.058
119	NEUROTECHNOLOGY-5	^{19}256	^{84}412	^{33}0.0042	^{33}0.0042	^{33}0.0042	^{54}1.026	^{31}0.0068	^{31}0.0068	^{31}0.0068	^{37}1.050
120	NEUROTECHNOLOGY-6	^{12}256	^{189}746	^{90}0.0153	^{90}0.0153	^{90}0.0153	^{83}1.070	^{95}0.0201	^{95}0.0201	^{95}0.0201	^{86}1.102
121	NEWLAND-2	^{116}2048	^{191}868				^{157}10.000	^{150}0.0811	^{150}0.0811	^{150}0.0811	^{153}1.491
122	NOBLE-1	^{140}2048	^{30}211	^{135}0.2049	^{135}0.2049	^{135}0.2049	^{135}2.390	^{181}0.2512	^{181}0.2512	^{181}0.2512	^{167}2.698
123	NOBLE-2	^{200}6144	^{110}535	^{132}0.1565	^{132}0.1565	^{132}0.1565	^{132}1.967	^{169}0.1816	^{169}0.1816	^{169}0.1816	^{166}2.098
124	NTECHLAB-0	^{196}4442	^{169}730	^{59}0.0077	^{59}0.0077	^{59}0.0077	^{53}1.038	^{63}0.0115	^{63}0.0115	^{63}0.0115	^{51}1.064
125	NTECHLAB-1	^{108}1736	^{82}405	^{69}0.0097	^{69}0.0097	^{69}0.0097	^{61}1.046	^{81}0.0139	^{81}0.0139	^{81}0.0139	^{69}1.074
126	NTECHLAB-3	^{174}3484	^{184}831	^{42}0.0051	^{42}0.0051	^{42}0.0051	^{23}1.024	^{41}0.0082	^{41}0.0082	^{41}0.0082	^{24}1.047
127	NTECHLAB-4	^{173}3484	^{198}929	^{31}0.0040	^{31}0.0040	^{31}0.0040	^{29}1.019	^{33}0.0068	^{33}0.0068	^{33}0.0068	^{21}1.041
128	NTECHLAB-5	^{108}1940	^{164}717	^{30}0.0039	^{30}0.0039	^{30}0.0039	^{27}1.018	^{28}0.0064	^{28}0.0064	^{28}0.0064	^{17}1.037
129	NTECHLAB-6	^{109}1940	^{187}841	^{26}0.0034	^{26}0.0034	^{26}0.0034	^{13}1.015	^{24}0.0059	^{24}0.0059	^{24}0.0059	^{16}1.034
130	QUANTASOFT-1	^{125}2048	^{76}396	^{149}0.9857	^{149}0.9857	^{149}0.9857	^{149}9.866	^{176}0.2198	^{176}0.2198	^{176}0.2198	^{180}2.559
131	RANKONE-0	^{72}228	^{6}50	^{108}0.0319	^{108}0.0319	^{108}0.0319	^{108}1.188	^{133}0.0455	^{133}0.0455	^{133}0.0455	^{131}1.275
132	RANKONE-1	^{27}324	^{17}136	^{97}0.0194	^{97}0.0194	^{97}0.0194	^{96}1.109	^{108}0.0247	^{108}0.0247	^{108}0.0247	^{108}1.145
133	RANKONE-2	^{11}133	^{14}113	^{89}0.0149	^{89}0.0149	^{89}0.0149	^{80}1.086	^{102}0.0221	^{102}0.0221	^{102}0.0221	^{102}1.135
134	RANKONE-3	^{9}133	^{16}114	^{88}0.0149	^{88}0.0149	^{88}0.0149	^{89}1.086	^{101}0.0221	^{101}0.0221	^{101}0.0221	^{101}1.135
135	RANKONE-4	^{1}85	^{4}36	^{107}0.0318	^{107}0.0318	^{107}0.0318	^{107}1.171	^{132}0.0441	^{132}0.0441	^{132}0.0441	^{136}1.249
136	RANKONE-5	^{10}133	^{13}94	^{58}0.0072	^{58}0.0072	^{58}0.0072	^{58}1.042	^{68}0.0120	^{68}0.0120	^{68}0.0120	^{69}1.078
137	REALNETWORKS-0	^{186}4100	^{38}244	^{115}0.0443	^{115}0.0443	^{115}0.0443	^{111}1.222	^{131}0.0426	^{131}0.0426	^{131}0.0426	^{122}1.222
138	REALNETWORKS-1	^{186}4104	^{37}243	^{111}0.0329	^{111}0.0329	^{111}0.0329	^{106}1.163	^{130}0.0426	^{130}0.0426	^{130}0.0426	^{123}1.222
139	REALNETWORKS-2	^{189}4104	^{39}245	^{109}0.0320	^{109}0.0320	^{109}0.0320	^{105}1.159	^{125}0.0418	^{125}0.0418	^{125}0.0418	^{120}1.217
140	REMARKAI-0	^{135}2048	^{138}615	^{54}0.0065	^{54}0.0065	^{54}0.0065	^{48}1.034	^{60}0.0109	^{60}0.0109	^{60}0.0109	^{58}1.065
141	REMARKAI-2	^{148}2048	^{58}434	^{51}0.0062	^{51}0.0062	^{51}0.0062	^{42}1.031	^{58}0.0105	^{58}0.0105	^{58}0.0105	^{48}1.061
142	SENSETIME-0	^{188}4104	^{192}715	^{13}0.0018	^{13}0.0018	^{13}0.0018	^{14}1.014	^{16}0.0048	^{16}0.0048	^{16}0.0048	^{33}1.040
143	SENSETIME-1	^{187}4104	^{142}656	^{11}0.0018	^{11}0.0018	^{11}0.0018	^{13}1.013	^{17}0.0048	^{17}0.0048	^{17}0.0048	^{32}1.040
144	SHAMAN-0	^{179}4096	^{113}538				^{181}10.000	^{168}0.1707	^{168}0.1707	^{168}0.1707	^{165}2.092

Table 17: **Rank-based accuracy for the FRVT 2018 mugshot sets.** *In columns 3 and 4 are template size and template generation duration. Thereafter values are rank-based FNIR with T = 0 and FPIR = 1. This is appropriate to investigational uses but not those with higher volumes where candidates from all searches would need review. Columns 5 - 9 show FRVT 2018 accuracy for various ranks for galleries unenrolled with all lifetime images. Column 10 is a workload statistic, a small value shows an algorithm front-loads mates into the first 10 candidates. The last four columns gives analogous results for enrollment only of the most recent image - see Figure 8. Throughout, blue superscripts indicate the rank of the algorithm for that column, and the best value is highlighted in yellow.*

MISSES OUTSIDE RANK R		RESOURCE USAGE		ENROLL LIFETIME CONSOLIDATED = 1.6M				ENROL MOST RECENT, N = 1.6M			
FNIR(N, T=0, R)		TEMPLATE		FRVT 2018 MUGSHOTS							
#	ALGORITHM	BYTES	MSEC	R=1	R=10	R=50	WORK-10	R=1	R=10	R=50	WORK-10
145	SHAMAN-1	^{157}4096	^{121}557				^{190}10.000	^{166}0.1718	^{166}0.1718	^{166}0.1718	^{184}2.078
146	SHAMAN-2	^{201}8192	^{122}557				^{194}10.000	^{182}0.2620	^{182}0.2620	^{182}0.2620	^{182}2.710
147	SHAMAN-3	^{121}2048	^{158}704	^{127}0.0969	^{127}0.0969	^{127}0.0969	^{128}1.613	^{155}0.1266	^{155}0.1266	^{155}0.1266	^{160}1.811
148	SHAMAN-4	^{130}2048	^{136}642	^{134}0.1867	^{134}0.1867	^{134}0.1867	^{134}2.163	^{178}0.2242	^{178}0.2242	^{178}0.2242	^{177}2.431
149	SHAMAN-6	^{145}2048	^{157}706	^{106}0.0312	^{106}0.0312	^{106}0.0312	^{114}1.249	^{129}0.0424	^{129}0.0424	^{129}0.0424	^{139}1.339
150	SHAMAN-7	^{174}2048	^{159}709	^{105}0.0310	^{105}0.0310	^{105}0.0310	^{113}1.248	^{128}0.0422	^{128}0.0422	^{128}0.0422	^{138}1.337
151	SIAT-0	^{88}1036	^{67}358				^{187}10.000	^{50}0.0101	^{50}0.0101	^{50}0.0101	^{47}1.059
152	SIAT-1	^{149}2052	^{138}842	^{138}0.2639	^{138}0.2639	^{138}0.2639	^{140}3.373	^{10}0.0039	^{10}0.0039	^{10}0.0039	^{17}1.031
153	SIAT-2	^{153}2052	^{195}906	^{136}0.2128	^{136}0.2128	^{136}0.2128	^{137}2.913	^{11}0.0040	^{11}0.0040	^{11}0.0040	^{22}1.032
154	SMILART-0	^{75}1024	^{30}168				^{191}10.000	^{170}0.1981	^{170}0.1981	^{170}0.1981	^{169}2.204
155	SMILART-1	^{85}1024	^{146}662				^{165}10.000	^{175}0.2188	^{175}0.2188	^{175}0.2188	^{176}2.435
156	SMILART-2	^{81}1024	^{123}560				^{155}10.000	^{171}0.1946	^{171}0.1946	^{171}0.1946	^{166}2.196
157	SMILART-4	^{38}512	^{18}167	^{147}0.9531	^{147}0.9531	^{147}0.9531	^{147}9.573	^{198}0.9649	^{198}0.9649	^{198}0.9649	^{198}9.679
158	SMILART-5	^{126}2048	^{103}464				^{178}10.000				^{201}10.000
159	SYNESIS-0	^{35}512	^{36}237				^{182}10.000	^{164}0.1621	^{164}0.1621	^{164}0.1621	^{173}2.380
160	SYNESIS-3	^{181}4096	^{12}103	^{131}0.1350	^{131}0.1350	^{131}0.1350	^{130}1.868	^{167}0.1721	^{167}0.1721	^{167}0.1721	^{167}2.140
161	TEVIAN-0	^{112}2048	^{75}394				^{153}10.000	^{104}0.0225	^{104}0.0225	^{104}0.0225	^{96}1.122
162	TEVIAN-1	^{146}2048	^{79}398				^{205}10.000	^{105}0.0225	^{105}0.0225	^{105}0.0225	^{100}1.122
163	TEVIAN-2	^{119}2048	^{77}397				^{167}10.000	^{103}0.0224	^{103}0.0224	^{103}0.0224	^{98}1.121
164	TEVIAN-3	^{129}2048	^{55}300	^{74}0.0102	^{74}0.0102	^{74}0.0102	^{67}1.052	^{88}0.0169	^{88}0.0169	^{88}0.0169	^{78}1.093
165	TEVIAN-4	^{136}2048	^{53}299	^{60}0.0080	^{60}0.0080	^{60}0.0080	^{56}1.041	^{75}0.0134	^{75}0.0134	^{75}0.0134	^{68}1.076
166	TEVIAN-5	^{132}2048	^{96}416	^{44}0.0053	^{44}0.0053	^{44}0.0053	^{39}1.028	^{48}0.0092	^{48}0.0092	^{48}0.0092	^{41}1.054
167	TIGER-0	^{155}2052	^{93}428	^{117}0.0480	^{117}0.0480	^{117}0.0480	^{112}1.247	^{144}0.0638	^{144}0.0638	^{144}0.0638	^{137}1.334
168	TIGER-1	^{156}2052	^{78}398				^{189}10.000				^{202}10.000
169	TIGER-2	^{152}2052	^{102}464	^{35}0.0044	^{35}0.0044	^{35}0.0044	^{31}1.023	^{39}0.0075	^{39}0.0075	^{39}0.0075	^{33}1.046
170	TIGER-3	^{147}2052	^{101}464				^{152}10.000	^{38}0.0075	^{38}0.0075	^{38}0.0075	^{33}1.046
171	TONGYITRANS-0	^{162}2070	^{27}190	^{49}0.0060	^{49}0.0060	^{49}0.0060	^{50}1.036	^{53}0.0095	^{53}0.0095	^{53}0.0095	^{50}1.062
172	TONGYITRANS-1	^{140}2070	^{25}189	^{90}0.0114	^{90}0.0114	^{90}0.0114	^{84}1.073	^{52}0.0095	^{52}0.0095	^{52}0.0095	^{51}1.062
173	TOSHIBA-0	^{93}1548	^{193}930	^{24}0.0033	^{24}0.0033	^{24}0.0033	^{23}1.018	^{32}0.0068	^{32}0.0068	^{32}0.0068	^{31}1.046
174	TOSHIBA-1	^{159}2060	^{201}931	^{28}0.0035	^{28}0.0035	^{28}0.0035	^{30}1.019	^{34}0.0071	^{34}0.0071	^{34}0.0071	^{36}1.047
175	VD-0	^{76}1028	^{81}337	^{143}0.4303	^{143}0.4303	^{143}0.4303	^{143}3.703	^{193}0.4751	^{192}0.4751	^{192}0.4751	^{191}4.074
176	VD-1	^{151}2052	^{158}695	^{102}0.0221	^{102}0.0221	^{102}0.0221	^{102}1.140	^{115}0.0302	^{115}0.0302	^{115}0.0302	^{119}1.197
177	VIGILANTSOLUTIONS-0	^{93}1544	^{180}823				^{166}10.000	^{154}0.1254	^{154}0.1254	^{154}0.1254	^{154}1.712
178	VIGILANTSOLUTIONS-1	^{156}2056	^{168}739				^{202}10.000	^{174}0.2038	^{174}0.2038	^{174}0.2038	^{170}2.210
179	VIGILANTSOLUTIONS-2	^{95}1544	^{177}820				^{177}10.000	^{180}0.2387	^{180}0.2387	^{180}0.2387	^{179}2.555
180	VIGILANTSOLUTIONS-3	^{97}1544	^{185}832	^{121}0.0549	^{121}0.0549	^{121}0.0549	^{116}1.280	^{148}0.0719	^{148}0.0719	^{148}0.0719	^{142}1.378
181	VIGILANTSOLUTIONS-4	^{92}1544	^{183}830	^{129}0.0993	^{129}0.0993	^{129}0.0993	^{124}1.549	^{156}0.1272	^{156}0.1272	^{156}0.1272	^{155}1.721
182	VIGILANTSOLUTIONS-5	^{94}1544	^{175}778				^{169}10.000	^{67}0.0118	^{67}0.0118	^{67}0.0118	^{60}1.069
183	VIGILANTSOLUTIONS-6	^{96}1544	^{186}834				^{181}10.000	^{70}0.0125	^{70}0.0125	^{70}0.0125	^{64}1.072
184	VISIONLABS-3	^{14}256	^{35}228	^{41}0.0050	^{41}0.0050	^{41}0.0050	^{57}1.041	^{46}0.0089	^{46}0.0089	^{46}0.0089	^{62}1.072
185	VISIONLABS-4	^{25}256	^{57}315	^{14}0.0020	^{14}0.0020	^{14}0.0020	^{12}1.013	^{13}0.0044	^{13}0.0044	^{13}0.0044	^{20}1.031
186	VISIONLABS-5	^{34}512	^{54}300	^{12}0.0018	^{12}0.0018	^{13}0.0018	^{11}1.012	^{12}0.0041	^{12}0.0041	^{12}0.0041	^{9}1.029
187	VISIONLABS-6	^{40}512	^{50}292	^{9}0.0015	^{9}0.0015	^{9}0.0015	^{10}1.011	^{7}0.0033	^{7}0.0033	^{7}0.0033	^{7}1.025
188	VISIONLABS-7	^{38}512	^{51}293	^{8}0.0014	^{8}0.0014	^{8}0.0014	^{8}1.010	^{6}0.0033	^{6}0.0033	^{6}0.0033	^{6}1.025
189	VOCORD-0	^{57}608	^{112}536				^{190}10.000	^{123}0.0403	^{123}0.0403	^{123}0.0403	^{136}1.301
190	VOCORD-1	^{56}608	^{111}536				^{180}10.000	^{122}0.0402	^{122}0.0402	^{122}0.0402	^{134}1.299
191	VOCORD-2	^{131}2048	^{134}635				^{187}10.000	^{120}0.0382	^{120}0.0382	^{120}0.0382	^{133}1.290
192	VOCORD-3	^{60}896	^{161}714	^{55}0.0067	^{55}0.0067	^{56}0.0067	^{54}1.039	^{43}0.0085	^{43}0.0085	^{43}0.0085	^{42}1.054
193	VOCORD-4	^{59}896	^{114}538	^{62}0.0084	^{62}0.0084	^{62}0.0084	^{66}1.051	^{57}0.0102	^{57}0.0102	^{57}0.0102	^{59}1.068
194	VOCORD-5	^{55}768	^{179}822	^{46}0.0057	^{46}0.0057	^{46}0.0057	^{51}1.036	^{49}0.0092	^{49}0.0092	^{49}0.0092	^{54}1.063
195	VOCORD-6	^{201}10240	^{181}825				^{201}10.000	^{203}1.0000	^{203}1.0000	^{203}1.0000	^{203}10.000
196	YISHENG-0	^{158}2108	^{137}615				^{158}10.000	^{111}0.0268	^{111}0.0268	^{111}0.0268	^{108}1.149
197	YISHENG-1	^{176}3704	^{74}387	^{99}0.0208	^{99}0.0208	^{99}0.0208	^{94}1.105	^{114}0.0290	^{114}0.0290	^{114}0.0290	^{109}1.156
198	YITU-0	^{191}4136	^{133}633	^{38}0.0047	^{38}0.0047	^{38}0.0047	^{43}1.031	^{36}0.0074	^{36}0.0074	^{36}0.0074	^{40}1.053
199	YITU-1	^{190}4136	^{200}930	^{36}0.0046	^{36}0.0046	^{36}0.0046	^{41}1.031	^{35}0.0072	^{35}0.0072	^{35}0.0072	^{38}1.052
200	YITU-2	^{192}4138	^{192}870	^{10}0.0015	^{10}0.0015	^{10}0.0015	^{7}1.010	^{14}0.0044	^{14}0.0044	^{14}0.0044	^{16}1.035
201	YITU-3	^{192}4138	^{193}871	^{17}0.0023	^{17}0.0023	^{17}0.0023	^{23}1.018	^{17}0.0054	^{17}0.0054	^{17}0.0054	^{25}1.044
202	YITU-4	^{193}2070	^{196}910	^{3}0.0011	^{3}0.0011	^{3}0.0011	^{5}1.008	^{5}0.0037	^{5}0.0037	^{5}0.0037	^{22}1.031
203	YITU-5	^{161}2070	^{190}861	^{18}0.0020	^{18}0.0020	^{18}0.0020	^{19}1.016	^{18}0.0048	^{18}0.0048	^{18}0.0048	^{28}1.041

Table 18: **Rank-based accuracy for the FRVT 2018 mugshot sets.** *In columns 3 and 4 are template size and template generation duration. Thereafter values are rank-based FNIR with T = 0 and FPIR = 1. This is appropriate to investigational uses but not those with higher volumes where candidates from all searches would need review. Columns 5 - 9 show FRVT 2018 accuracy for various ranks for galleries unenrolled with all lifetime images. Column 10 is a workload statistic, a small value shows an algorithm front-loads mates into the first 10 candidates. The last four columns gives analogous results for enrollment only of the most recent image - see Figure 8. Throughout, blue superscripts indicate the rank of the algorithm for that column, and the best value is highlighted in yellow.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

MISSES BELOW THRESHOLD, T FNIR(N, T > 0, R > L)		ENROL MOST RECENT MUGSHOT, N = 1.6M DATASET: FRVT 2018 MUGSHOTS			DATASET: WEBCAM PROBES			DATASET: PROFILE PROBES		
#	ALGORITHM	FPIR=0.001	FPIR=0.01	FPIR=0.1	FPIR=0.001	FPIR=0.01	FPIR=0.1	FPIR=0.001	FPIR=0.01	FPIR=0.1
1	3DIVI-0	^{128}0.256	^{134}0.160	^{135}0.086	^{115}0.425	^{117}0.302	^{115}0.180			
2	3DIVI-1	^{125}0.256	^{135}0.160	^{136}0.087						
3	3DIVI-2	^{121}0.255	^{136}0.164	^{137}0.089						
4	3DIVI-3	^{145}0.402	^{152}0.284	^{152}0.168	^{131}0.626	^{133}0.497	^{139}0.243			
5	3DIVI-4	^{106}0.171	^{107}0.096	^{101}0.047	^{108}0.348	^{108}0.237	^{109}0.138			
6	3DIVI-5	^{101}0.169	^{106}0.095	^{102}0.047	^{106}0.339	^{107}0.234	^{108}0.137	^{30}0.995	^{25}0.987	^{28}0.960
7	3DIVI-6	^{104}0.170	^{110}0.098	^{107}0.051	^{107}0.342	^{109}0.238	^{110}0.142			
8	ALCHERA-0	^{96}0.140	^{94}0.073	^{91}0.035	^{77}0.218	^{78}0.146	^{79}0.087			
9	ALCHERA-1	^{196}0.999	^{198}0.999	^{199}0.995	^{164}1.000	^{169}1.000	^{161}1.000			
10	ALCHERA-2	^{156}0.490	^{153}0.304	^{154}0.184	^{128}0.591	^{127}0.442	^{136}0.295			
11	ALCHERA-3	^{98}0.159	^{95}0.073	^{84}0.030	^{84}0.239	^{82}0.152	^{79}0.081	^{28}0.999	^{27}0.993	^{22}0.921
12	ANKE-0	^{83}0.120	^{85}0.065	^{87}0.033	^{79}0.220	^{80}0.151	^{82}0.088	^{18}0.991	^{24}0.985	^{25}0.972
13	ANKE-1	^{89}0.122	^{84}0.065	^{88}0.033	^{78}0.220	^{81}0.151	^{81}0.088			
14	AWARE-0	^{194}0.983	^{126}0.128	^{138}0.085	^{143}0.817	^{111}0.253	^{114}0.178			
15	AWARE-1	^{195}0.996	^{125}0.127	^{132}0.081						
16	AWARE-2	^{192}0.977	^{122}0.120	^{130}0.078						
17	AWARE-3	^{93}0.131	^{100}0.085	^{108}0.051	^{99}0.298	^{100}0.204	^{107}0.132			
18	AWARE-4	^{127}0.271	^{139}0.177	^{134}0.107	^{123}0.509	^{125}0.375	^{134}0.253			
19	AWARE-5	^{139}0.373	^{103}0.088	^{105}0.050	^{87}0.253	^{85}0.163	^{86}0.099	^{30}1.000	^{53}0.999	^{33}0.998
20	AWARE-6	^{128}0.278	^{140}0.178	^{146}0.109	^{112}0.398	^{115}0.283	^{116}0.188			
21	AYONIX-0	^{180}0.811	^{187}0.725	^{190}0.598	^{153}0.939	^{154}0.892	^{156}0.802			
22	AYONIX-1	^{183}0.825	^{185}0.702	^{183}0.526	^{148}0.920	^{150}0.845	^{151}0.703			
23	AYONIX-2	^{182}0.825	^{186}0.702	^{187}0.526	^{149}0.920	^{149}0.845	^{150}0.702			
24	CAMVI-1	^{175}0.684	^{176}0.549	^{178}0.375	^{140}0.770	^{144}0.648	^{144}0.488			
25	CAMVI-2	^{160}0.537	^{164}0.402	^{161}0.242						
26	CAMVI-3	^{56}0.074	^{82}0.060	^{115}0.055	^{90}0.182	^{69}0.108	^{88}0.094			
27	CAMVI-4	^{57}0.074	^{79}0.056	^{104}0.050	^{68}0.136	^{58}0.100	^{75}0.083	^{26}0.999	^{31}0.994	^{15}0.816
28	CAMVI-5	^{72}0.102	^{98}0.078	^{123}0.069	^{73}0.179	^{78}0.132	^{98}0.110			
29	COGENT-0	^{45}0.056	^{52}0.032	^{61}0.020	^{57}0.140	^{62}0.100	^{71}0.069			
30	COGENT-1	^{44}0.056	^{51}0.032	^{60}0.020	^{60}0.140	^{61}0.100	^{70}0.069			
31	COGENT-2	^{30}0.047	^{19}0.020	^{21}0.010	^{24}0.098	^{28}0.063	^{26}0.036	^{22}0.997	^{26}0.993	^{31}0.982
32	COGENT-3	^{36}0.051	^{18}0.018	^{19}0.009	^{20}0.095	^{23}0.061	^{27}0.037			
33	COGNITEC-0	^{99}0.163	^{108}0.098	^{111}0.053	^{100}0.303	^{98}0.200	^{97}0.115			
34	COGNITEC-1	^{77}0.105	^{77}0.055	^{78}0.027	^{82}0.230	^{77}0.135	^{72}0.071			
35	COGNITEC-2	^{46}0.056	^{41}0.027	^{39}0.014	^{72}0.178	^{64}0.101	^{59}0.050	^{31}1.000	^{16}0.947	^{23}0.936
36	COGNITEC-3	^{43}0.055	^{44}0.028	^{41}0.014	^{50}0.162	^{59}0.100	^{61}0.050			
37	DAHUA-0	^{67}0.089	^{74}0.047	^{68}0.022	^{47}0.135	^{47}0.083	^{46}0.046			
38	DAHUA-1	^{59}0.075	^{58}0.039	^{53}0.018	^{41}0.122	^{40}0.075	^{39}0.042	^{10}0.953	^{10}0.862	^{15}0.679
39	DERMALOG-0	^{156}0.488	^{159}0.364	^{160}0.233	^{136}0.657	^{139}0.528	^{139}0.362			
40	DERMALOG-1	^{158}0.528	^{165}0.405	^{165}0.268						
41	DERMALOG-2	^{157}0.503	^{161}0.378	^{162}0.244						
42	DERMALOG-3	^{154}0.484	^{158}0.362	^{158}0.231	^{133}0.655	^{138}0.526	^{135}0.361			
43	DERMALOG-4	^{155}0.481	^{157}0.360	^{157}0.230	^{134}0.657	^{136}0.526	^{132}0.359			
44	DERMALOG-5	^{71}0.091	^{63}0.045	^{74}0.024	^{67}0.154	^{56}0.096	^{66}0.057			
45	DERMALOG-6	^{41}0.054	^{45}0.028	^{44}0.015	^{27}0.105	^{29}0.067	^{30}0.039	^{8}0.948	^{9}0.856	^{10}0.642
46	EVERAI-0	^{73}0.092	^{73}0.047	^{80}0.028	^{67}0.170	^{60}0.100	^{64}0.060			
47	EVERAI-1	^{37}0.052	^{27}0.023	^{22}0.010	^{42}0.128	^{36}0.074	^{32}0.039			
48	EVERAI-2	^{38}0.053	^{38}0.025	^{27}0.011	^{40}0.119	^{41}0.076	^{38}0.041			
49	EVERAI-3	^{17}0.038	^{17}0.018	^{17}0.008	^{21}0.096	^{21}0.060	^{22}0.034	^{14}0.979	^{8}0.535	^{9}0.247
50	EYEDEA-0	^{181}0.812	^{184}0.679	^{184}0.484	^{147}0.914	^{147}0.783	^{149}0.619			
51	EYEDEA-1	^{168}0.632	^{169}0.480	^{172}0.365						
52	EYEDEA-2	^{176}0.794	^{173}0.490	^{174}0.338						
53	EYEDEA-3	^{142}0.389	^{150}0.267	^{150}0.160	^{125}0.543	^{126}0.404	^{135}0.264			
54	GLORY-0	^{135}0.369	^{154}0.297	^{159}0.233	^{126}0.547	^{130}0.470	^{138}0.290			
55	GLORY-1	^{138}0.307	^{147}0.238	^{158}0.179	^{124}0.537	^{128}0.448	^{130}0.352			
56	GORILLA-0									
57	GORILLA-1	^{146}0.408	^{148}0.248	^{147}0.136	^{112}0.453	^{119}0.314	^{119}0.191			
58	GORILLA-2	^{108}0.190	^{114}0.108	^{106}0.051	^{92}0.268	^{90}0.170	^{88}0.093			
59	GORILLA-3	^{128}0.326	^{132}0.160	^{124}0.074	^{117}0.434	^{110}0.247	^{105}0.131			
60	HBINNO-0	^{177}0.766	^{182}0.632	^{183}0.458						
61	HIK-0	^{82}0.114	^{93}0.070	^{95}0.040	^{68}0.155	^{66}0.103	^{66}0.061			
62	HIK-1	^{87}0.120	^{91}0.067	^{90}0.034						
63	HIK-2	^{88}0.121	^{92}0.067	^{89}0.034						
64	HIK-3	^{76}0.105	^{82}0.060	^{85}0.030	^{70}0.158	^{67}0.105	^{63}0.061			
65	HIK-4	^{74}0.101	^{80}0.056	^{82}0.029	^{58}0.153	^{63}0.101	^{59}0.059			
66	HIK-5	^{27}0.047	^{23}0.022	^{29}0.011	^{11}0.077	^{11}0.048	^{15}0.028	^{27}0.999	^{29}0.994	^{12}0.662
67	HIK-6	^{32}0.050	^{26}0.022	^{38}0.011	^{12}0.086	^{14}0.052	^{16}0.029	^{32}1.000	^{82}0.997	^{11}0.645
68	IDEMIA-0	^{81}0.114	^{85}0.062	^{81}0.029	^{85}0.240	^{83}0.156	^{78}0.085			
69	IDEMIA-1	^{40}0.054	^{50}0.031	^{54}0.018						
70	IDEMIA-2	^{42}0.054	^{53}0.032	^{56}0.019						
71	IDEMIA-3	^{31}0.050	^{37}0.024	^{40}0.014	^{56}0.165	^{43}0.079	^{54}0.050			
72	IDEMIA-4	^{19}0.040	^{31}0.024	^{42}0.014	^{39}0.118	^{45}0.079	^{52}0.050	^{11}0.969	^{17}0.962	^{30}0.952

Table 19: **Threshold-based accuracy.** *Values are FNIR(N, T, L) with N = 1.6 million with thresholds set to produce FPIR = 0.001, 0.01, and 0.1 in non-mate searches. Columns 3-5 apply to FRVT-2018 mugshots: Columns 6-8 show the corresponding FNIR values for webcam images searched against the FRVT-2018 mugshot gallery. Finally, the three rightmost columns show FNIR for profile view images searched against the FRVT-2018 frontal gallery. Throughout blue superscripts indicate the rank of the algorithm for that column. Caution: The Power-low models are mostly intended to draw attention to the kind of behavior, not as a model to be used for prediction.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

MISSES BELOW THRESHOLD, T FNIR(N, T > 0, R > L)		ENROL MOST RECENT MUGSHOT, N = 1.6M								
		DATASET: FRVT 2018 MUGSHOTS			DATASET: WEBCAM PROBES			DATASET: PROFILE PROBES		
#	ALGORITHM	FPIR=0.001	FPIR=0.01	FPIR=0.1	FPIR=0.001	FPIR=0.01	FPIR=0.1	FPIR=0.001	FPIR=0.01	FPIR=0.1
73	IDEMIA-5	^{26}0.047	^{46}0.028	^{49}0.017	^{55}0.150	^{65}0.102	^{69}0.065	^{12}0.974	^{19}0.968	^{27}0.960
74	IDEMIA-6	^{24}0.046	^{43}0.028	^{51}0.018	^{81}0.226	^{89}0.161	^{95}0.108			
75	IMAGUS-0	^{175}0.784	^{181}0.608	^{182}0.453	^{145}0.872	^{146}0.779	^{148}0.635			
76	IMAGUS-2	^{176}0.751	^{179}0.566	^{177}0.377	^{142}0.816	^{143}0.645	^{142}0.460			
77	IMAGUS-3	^{179}0.808	^{183}0.670	^{186}0.512	^{146}0.909	^{146}0.809	^{149}0.667			
78	INCODE-0	^{134}0.313	^{144}0.201	^{143}0.107	^{114}0.420	^{118}0.304	^{117}0.191			
79	INCODE-1	^{114}0.214	^{115}0.114	^{103}0.060	^{96}0.296	^{96}0.198	^{96}0.110			
80	INCODE-2	^{107}0.186	^{112}0.102	^{100}0.046	^{93}0.269	^{91}0.176	^{87}0.100			
81	INCODE-3	^{103}0.170	^{101}0.086	^{94}0.037	^{90}0.264	^{87}0.164	^{80}0.087			
82	INNOVATRICS-0	^{124}0.255	^{138}0.165	^{139}0.089	^{109}0.361	^{112}0.258	^{113}0.159			
83	INNOVATRICS-1	^{123}0.255	^{137}0.165	^{138}0.089						
84	INNOVATRICS-2	^{120}0.237	^{132}0.142	^{131}0.079	^{101}0.310	^{102}0.209	^{101}0.126			
85	INNOVATRICS-3	^{116}0.224	^{128}0.134	^{122}0.068	^{97}0.297	^{79}0.203	^{93}0.116			
86	INNOVATRICS-4	^{94}0.134	^{82}0.076	^{93}0.035	^{80}0.222	^{80}0.149	^{77}0.085	^{13}0.977	^{18}0.966	^{24}0.945
87	ISYSTEMS-0	^{72}0.091	^{69}0.047	^{72}0.023	^{69}0.173	^{80}0.110	^{67}0.065			
88	ISYSTEMS-1	^{69}0.090	^{67}0.047	^{71}0.023						
89	ISYSTEMS-2	^{62}0.081	^{53}0.035	^{46}0.015	^{42}0.126	^{45}0.080	^{46}0.046			
90	ISYSTEMS-3	^{52}0.062	^{40}0.027	^{36}0.012	^{30}0.107	^{31}0.068	^{31}0.039	^{29}1.000	^{31}0.995	^{21}0.913
91	LOOKMAN-3	^{25}0.046	^{41}0.027	^{50}0.017	^{33}0.112	^{46}0.082	^{57}0.057			
92	LOOKMAN-4	^{28}0.047	^{39}0.027	^{47}0.016	^{29}0.105	^{31}0.075	^{56}0.052	^{16}0.980	^{22}0.978	^{30}0.977
93	MEGVII-0	^{80}0.109	^{51}0.058	^{77}0.025	^{35}0.116	^{30}0.067	^{33}0.034			
94	MEGVII-1	^{59}0.075	^{57}0.039	^{67}0.022	^{23}0.097	^{22}0.061	^{21}0.033			
95	MEGVII-2	^{61}0.080	^{59}0.039	^{65}0.022	^{22}0.096	^{21}0.059	^{20}0.033	^{23}0.997	^{9}0.698	^{7}0.429
96	MICROFOCUS-0	^{188}0.933	^{192}0.867	^{194}0.749	^{158}0.985	^{157}0.950	^{156}0.877			
97	MICROFOCUS-1	^{189}0.933	^{193}0.867	^{195}0.749						
98	MICROFOCUS-2	^{190}0.934	^{194}0.870	^{196}0.758						
99	MICROFOCUS-3	^{187}0.931	^{191}0.866	^{193}0.748	^{157}0.979	^{156}0.948	^{157}0.876			
100	MICROFOCUS-4	^{197}0.999	^{199}0.999	^{198}0.994	^{155}0.975	^{155}0.940	^{155}0.862			
101	MICROFOCUS-5	^{184}0.836	^{189}0.736	^{189}0.588	^{151}0.928	^{152}0.865	^{151}0.748			
102	MICROFOCUS-6	^{193}0.978	^{196}0.963	^{191}0.641	^{150}0.923	^{151}0.858	^{150}0.739			
103	MICROSOFT-0	^{21}0.044	^{22}0.022	^{28}0.010	^{36}0.115	^{33}0.071	^{31}0.040			
104	MICROSOFT-1	^{23}0.045	^{24}0.022	^{29}0.011						
105	MICROSOFT-2	^{34}0.050	^{36}0.026	^{34}0.012						
106	MICROSOFT-3	^{16}0.030	^{16}0.014	^{12}0.006	^{17}0.091	^{19}0.056	^{14}0.028			
107	MICROSOFT-4	^{13}0.029	^{13}0.013	^{10}0.005	^{13}0.087	^{15}0.053	^{13}0.026			
108	MICROSOFT-5	^{12}0.028	^{12}0.012	^{7}0.005	^{10}0.070	^{7}0.041	^{7}0.021	^{2}0.338	^{2}0.188	^{2}0.123
109	MICROSOFT-6	^{5}0.014	^{5}0.008	^{3}0.004	^{5}0.037	^{5}0.024	^{4}0.016	^{1}0.203	^{1}0.148	^{1}0.109
110	NEC-0	^{83}0.082	^{74}0.049	^{83}0.029	^{52}0.140	^{52}0.093	^{60}0.059			
111	NEC-1	^{79}0.108	^{87}0.063	^{92}0.035	^{78}0.197	^{76}0.138	^{76}0.083			
112	NEC-2	^{2}0.005	^{1}0.004	^{1}0.003	^{2}0.020	^{2}0.013	^{1}0.010			
113	NEC-3	^{1}0.004	^{2}0.004	^{2}0.003	^{1}0.017	^{1}0.013	^{2}0.011	^{5}0.664	^{5}0.479	^{6}0.340
114	NEUROTECHNOLOGY-0	^{129}0.295	^{142}0.196	^{145}0.108	^{119}0.465	^{130}0.317	^{120}0.196			
115	NEUROTECHNOLOGY-1	^{131}0.299	^{142}0.195	^{142}0.105						
116	NEUROTECHNOLOGY-2	^{132}0.299	^{141}0.195	^{140}0.105						
117	NEUROTECHNOLOGY-3	^{172}0.665	^{111}0.101	^{110}0.052	^{91}0.266	^{88}0.164	^{83}0.088			
118	NEUROTECHNOLOGY-4	^{54}0.066	^{48}0.030	^{43}0.014	^{36}0.117	^{38}0.073	^{30}0.040			
119	NEUROTECHNOLOGY-5	^{48}0.056	^{37}0.025	^{33}0.012	^{44}0.130	^{37}0.074	^{40}0.042	^{21}0.996	^{26}0.982	^{25}0.948
120	NEUROTECHNOLOGY-6	^{127}0.255	^{123}0.124	^{109}0.051	^{113}0.418	^{101}0.206	^{88}0.103			
121	NEWLAND-2	^{150}0.441	^{154}0.296	^{149}0.157	^{120}0.466	^{122}0.335	^{122}0.213			
122	NOBLIS-1	^{199}1.000	^{187}0.992	^{180}0.419	^{159}1.000	^{159}1.000	^{160}1.000			
123	NOBLIS-2	^{196}0.997	^{173}0.490	^{169}0.309	^{168}1.000	^{158}1.000	^{146}0.565	^{33}1.000	^{30}1.000	^{34}1.000
124	NTECHLAB-0	^{64}0.083	^{72}0.047	^{69}0.023	^{64}0.162	^{48}0.105	^{62}0.061			
125	NTECHLAB-1	^{76}0.102	^{78}0.056	^{79}0.027						
126	NTECHLAB-3	^{47}0.056	^{49}0.030	^{45}0.015	^{37}0.118	^{30}0.075	^{41}0.043			
127	NTECHLAB-4	^{30}0.043	^{29}0.024	^{32}0.012	^{28}0.105	^{28}0.065	^{27}0.036			
128	NTECHLAB-5	^{27}0.045	^{30}0.024	^{30}0.012	^{36}0.102	^{35}0.063	^{25}0.034			
129	NTECHLAB-6	^{18}0.039	^{20}0.021	^{23}0.010	^{19}0.094	^{19}0.059	^{18}0.032	^{3}0.566	^{3}0.443	^{3}0.317
130	QUANTASOFT-1	^{170}0.640	^{172}0.494	^{173}0.335						
131	RANKONE-0	^{115}0.219	^{127}0.129	^{129}0.078	^{111}0.391	^{116}0.291	^{119}0.195			
132	RANKONE-1	^{100}0.168	^{107}0.087	^{95}0.043						
133	RANKONE-2	^{38}0.120	^{87}0.073	^{97}0.042	^{89}0.261	^{96}0.190	^{100}0.126			
134	RANKONE-3	^{84}0.120	^{76}0.073	^{96}0.042	^{88}0.255	^{92}0.187	^{97}0.122			
135	RANKONE-4	^{109}0.195	^{124}0.126	^{125}0.076	^{116}0.426	^{123}0.324	^{123}0.221			
136	RANKONE-5	^{50}0.062	^{56}0.036	^{62}0.021	^{70}0.173	^{73}0.119	^{73}0.074	^{20}0.998	^{23}0.994	^{32}0.988
137	REALNETWORKS-0	^{119}0.236	^{131}0.140	^{128}0.077	^{104}0.319	^{106}0.209	^{109}0.129			
138	REALNETWORKS-1	^{118}0.236	^{130}0.140	^{127}0.077	^{103}0.319	^{103}0.209	^{102}0.129			
139	REALNETWORKS-2	^{117}0.234	^{129}0.139	^{126}0.077	^{102}0.318	^{106}0.209	^{104}0.129			
140	REMARKAI-0	^{92}0.130	^{86}0.062	^{79}0.025	^{76}0.203	^{79}0.123	^{66}0.064			
141	REMARKAI-2	^{91}0.126	^{84}0.061	^{78}0.024	^{74}0.196	^{72}0.122	^{65}0.063	^{15}0.980	^{16}0.958	^{18}0.878
142	SENSETIME-0	^{9}0.023	^{10}0.012	^{14}0.007	^{8}0.063	^{8}0.040	^{7}0.025	^{124}1.000	^{20}0.971	^{16}0.844
143	SENSETIME-1	^{11}0.025	^{11}0.012	^{15}0.007	^{9}0.064	^{10}0.041	^{11}0.025			
144	SHAMAN-0	^{152}0.474	^{160}0.370	^{164}0.259	^{130}0.621	^{124}0.507	^{138}0.375			

Table 20: **Threshold-based accuracy.** *Values are FNIR(N, T, L) with N = 1.6 million with thresholds set to produce FPIR = 0.001, 0.01, and 0.1 in non-mate searches. Columns 3-5 apply to FRVT-2018 mugshots: Columns 6-8 show the corresponding FNIR values for webcam images searched against the FRVT-2018 mugshot gallery. Finally, the three rightmost columns show FNIR for profile view images searched against the FRVT-2018 frontal gallery. Throughout blue superscripts indicate the rank of the algorithm for that column. Caution: The Power-low models are mostly intended to draw attention to the kind of behavior, not as a model to be used for prediction.*

MISSES BELOW THRESHOLD, T		ENROL MOST RECENT MUGSHOT, N = 1.6M								
FNIR(N, T > 0, R > 1)		DATASET: FRVT 2018 MUGSHOTS			DATASET: WEBCAM PROBES			DATASET: PROFILE PROBES		
#	ALGORITHM	FPIR=0.001	FPIR=0.01	FPIR=0.1	FPIR=0.001	FPIR=0.01	FPIR=0.1	FPIR=0.001	FPIR=0.01	FPIR=0.1
145	SHAMAN-1	^{156}0.532	^{166}0.406	^{157}0.274						
146	SHAMAN-2	^{174}0.700	^{180}0.582	^{165}0.424						
147	SHAMAN-3	^{151}0.453	^{156}0.348	^{150}0.225	^{139}0.597	^{131}0.472	^{127}0.317			
148	SHAMAN-4	^{165}0.616	^{171}0.490	^{170}0.344	^{139}0.754	^{142}0.639	^{143}0.480			
149	SHAMAN-6	^{98}0.143	^{105}0.095	^{119}0.060	^{83}0.237	^{88}0.168	^{92}0.108	^{9}0.952	^{16}0.935	^{20}0.905
150	SHAMAN-7	^{97}0.144	^{104}0.094	^{118}0.060	^{86}0.240	^{89}0.169	^{90}0.107			
151	SIAT-0	^{70}0.091	^{68}0.047	^{64}0.022	^{31}0.107	^{27}0.064	^{28}0.025			
152	SIAT-1	^{6}0.020	^{6}0.009	^{8}0.005	^{110}0.365	^{123}0.348	^{126}0.337			
153	SIAT-2	^{10}0.024	^{7}0.009	^{6}0.005	^{124}0.478	^{129}0.460	^{140}0.451			
154	SMILART-0	^{166}0.620	^{170}0.486	^{170}0.322						
155	SMILART-1	^{171}0.641	^{177}0.505	^{178}0.342						
156	SMILART-2	^{167}0.629	^{174}0.492	^{171}0.325						
157	SMILART-4	^{191}0.968	^{196}0.965	^{197}0.964	^{156}0.976	^{188}0.973	^{159}0.973			
158	SMILART-5									
159	SYNESIS-0	^{163}0.554	^{162}0.378	^{150}0.218	^{138}0.734	^{141}0.598	^{140}0.431			
160	SYNESIS-3	^{164}0.583	^{168}0.444	^{166}0.294	^{132}0.646	^{136}0.524	^{138}0.372			
161	TEVIAN-0	^{111}0.203	^{117}0.114	^{113}0.054	^{103}0.331	^{106}0.227	^{116}0.132			
162	TEVIAN-1	^{112}0.203	^{118}0.114	^{114}0.054						
163	TEVIAN-2	^{110}0.202	^{116}0.114	^{112}0.054						
164	TEVIAN-3	^{106}0.180	^{109}0.098	^{99}0.044	^{98}0.298	^{97}0.198	^{98}0.113			
165	TEVIAN-4	^{88}0.120	^{90}0.066	^{88}0.031	^{71}0.176	^{73}0.115	^{82}0.065			
166	TEVIAN-5	^{68}0.090	^{71}0.047	^{66}0.022	^{53}0.144	^{49}0.089	^{51}0.049	^{7}0.910	^{7}0.661	^{5}0.483
167	TIGER-0	^{143}0.392	^{152}0.262	^{148}0.142	^{122}0.500	^{124}0.366	^{121}0.211			
168	TIGER-1				^{127}0.580	^{132}0.487	^{139}0.396			
169	TIGER-2	^{66}0.089	^{61}0.042	^{58}0.018	^{62}0.158	^{55}0.095	^{40}0.048	^{28}0.998	^{12}0.927	^{9}0.503
170	TIGER-3	^{65}0.089	^{62}0.042	^{57}0.018	^{61}0.158	^{54}0.095	^{40}0.048			
171	TONGYITRANS-0	^{60}0.077	^{60}0.041	^{67}0.019	^{32}0.112	^{32}0.069	^{30}0.038			
172	TONGYITRANS-1	^{56}0.069	^{54}0.035	^{48}0.016	^{25}0.101	^{24}0.062	^{34}0.034			
173	TOSHIBA-0	^{53}0.065	^{47}0.029	^{37}0.013	^{35}0.118	^{35}0.074	^{38}0.041	^{17}0.988	^{22}0.971	^{25}0.899
174	TOSHIBA-1	^{51}0.062	^{21}0.021	^{24}0.010	^{18}0.092	^{16}0.054	^{19}0.032			
175	VD-0	^{188}0.917	^{190}0.828	^{192}0.668	^{173}0.946	^{163}0.871	^{172}0.725			
176	VD-1	^{115}0.204	^{121}0.118	^{117}0.059	^{84}0.281	^{96}0.188	^{89}0.106			
177	VIGILANTSOLUTIONS-0	^{161}0.539	^{163}0.304	^{158}0.247	^{137}0.695	^{140}0.557	^{137}0.389			
178	VIGILANTSOLUTIONS-1	^{168}0.637	^{176}0.502	^{177}0.348						
179	VIGILANTSOLUTIONS-2	^{185}0.876	^{188}0.731	^{186}0.489						
180	VIGILANTSOLUTIONS-3	^{147}0.410	^{151}0.283	^{151}0.163	^{136}0.660	^{137}0.526	^{130}0.356			
181	VIGILANTSOLUTIONS-4	^{162}0.550	^{167}0.424	^{156}0.268	^{144}0.817	^{145}0.709	^{145}0.623			
182	VIGILANTSOLUTIONS-5	^{149}0.433	^{63}0.045	^{80}0.023						
183	VIGILANTSOLUTIONS-6	^{148}0.426	^{66}0.046	^{73}0.023						
184	VISIONLABS-3	^{36}0.051	^{37}0.026	^{38}0.013	^{69}0.187	^{81}0.091	^{55}0.051			
185	VISIONLABS-4	^{49}0.060	^{38}0.026	^{20}0.010	^{63}0.159	^{57}0.097	^{42}0.045			
186	VISIONLABS-5	^{39}0.053	^{25}0.022	^{15}0.008	^{54}0.147	^{48}0.087	^{37}0.041			
187	VISIONLABS-6	^{15}0.029	^{14}0.012	^{11}0.005	^{16}0.090	^{13}0.051	^{12}0.025			
188	VISIONLABS-7	^{14}0.029	^{13}0.012	^{9}0.005	^{15}0.090	^{12}0.051	^{10}0.025	^{2}0.461	^{3}0.322	^{3}0.198
189	VOCORD-0	^{144}0.399	^{120}0.116	^{121}0.062	^{85}0.285	^{92}0.181	^{91}0.108			
190	VOCORD-1	^{130}0.299	^{119}0.116	^{120}0.062						
191	VOCORD-2	^{137}0.366	^{115}0.107	^{116}0.057						
192	VOCORD-3	^{90}0.126	^{75}0.050	^{59}0.020	^{59}0.155	^{53}0.093	^{45}0.048			
193	VOCORD-4	^{140}0.378	^{76}0.054	^{63}0.021	^{68}0.173	^{51}0.093	^{44}0.046			
194	VOCORD-5	^{112}0.170	^{66}0.046	^{55}0.019	^{45}0.130	^{44}0.080	^{42}0.043	^{19}0.992	^{13}0.929	^{19}0.787
195	VOCORD-6	^{203}1.000	^{203}1.000	^{203}1.000	^{201}1.000	^{201}1.000	^{201}1.000			
196	YISHENG-0	^{141}0.380	^{146}0.209	^{134}0.086	^{154}0.974	^{114}0.275	^{112}0.146			
197	YISHENG-1	^{136}0.348	^{145}0.208	^{140}0.090	^{141}0.808	^{113}0.269	^{110}0.144			
198	YITU-0	^{38}0.050	^{33}0.025	^{35}0.012	^{14}0.090	^{17}0.054	^{17}0.030			
199	YITU-1	^{29}0.047	^{28}0.023	^{30}0.011						
200	YITU-2	^{7}0.020	^{8}0.011	^{13}0.006	^{6}0.049	^{6}0.028	^{5}0.016			
201	YITU-3	^{8}0.021	^{7}0.011	^{16}0.007	^{7}0.052	^{7}0.033	^{7}0.021			
202	YITU-4	^{3}0.012	^{3}0.007	^{4}0.004	^{3}0.027	^{3}0.017	^{3}0.011	^{8}0.902	^{17}0.975	^{37}0.845
203	YITU-5	^{4}0.013	^{4}0.007	^{5}0.005	^{4}0.032	^{4}0.023	^{6}0.017			

Table 21: **Threshold-based accuracy.** *Values are FNIR(N, T, L) with N = 1.6 million with thresholds set to produce FPIR = 0.001, 0.01, and 0.1 in non-mate searches. Columns 3-5 apply to FRVT-2018 mugshots: Columns 6-8 show the corresponding FNIR values for webcam images searched against the FRVT-2018 mugshot gallery. Finally, the three rightmost columns show FNIR for profile view images searched against the FRVT-2018 frontal gallery. Throughout blue superscripts indicate the rank of the algorithm for that column. Caution: The Power-low models are mostly intended to draw attention to the kind of behavior, not as a model to be used for prediction.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

		INVESTIGATION MODE				IDENTIFICATION MODE				FAILURE TO EXTRACT			
		RANK ONE MISS RATE, FNIR(N, 0, 1)				HIGH T → FPIR = 0.01, FNIR(N, T, 1)				FEATURES			
#	ALGORITHM	N=1.6M FRVT-18	N=1.6M WEBCAM	N=1.6M PROFILE	N=1.1M WILD	N=1.6M FRVT-18	N=1.6M WEBCAM	N=1.6M PROFILE	N=1.1M WILD+	N=1.6M FRVT-18	N=1.6M WEBCAM	N=1.6M PROFILE	N=1.1M WILD
1	3DIVI-0	^{118}0.034	^{111}0.086		^{59}0.071	^{136}0.160	^{117}0.302		^{60}0.095	0.003	0.007		0.013
2	3DIVI-1	^{119}0.038			^{62}0.074	^{138}0.160			^{65}0.095	0.002			0.013
3	3DIVI-2	^{124}0.040			^{64}0.076	^{139}0.164			^{62}0.096	0.003			0.013
4	3DIVI-3	^{152}0.086	^{128}0.206		^{71}0.094	^{152}0.284	^{133}0.497		^{85}0.136	0.002	0.005		0.009
5	3DIVI-4	^{96}0.020	^{98}0.062			^{107}0.096	^{108}0.237			0.002	0.005		
6	3DIVI-5	^{97}0.020	^{96}0.062	^{23}0.894	^{32}0.052	^{106}0.095	^{107}0.234	^{25}0.987	^{42}0.069	0.002	0.005	0.442	0.004
7	3DIVI-6	^{110}0.027	^{106}0.074		^{38}0.060	^{110}0.098	^{109}0.238		^{46}0.072	0.002	0.005		0.004
8	ALCHERA-0	^{92}0.019	^{88}0.047		^{77}0.092	^{58}0.073	^{78}0.146		^{75}0.089	0.006	0.014		0.030
9	ALCHERA-1	^{199}0.987	^{169}1.000			^{198}0.999	^{169}1.000			0.006	0.013		
10	ALCHERA-2	^{153}0.097	^{135}0.166		^{84}0.098	^{159}0.304	^{127}0.442		^{88}0.135	0.001	0.002		0.012
11	ALCHERA-3	^{72}0.013	^{68}0.035	^{16}0.629	^{36}0.064	^{56}0.073	^{82}0.152	^{27}0.998	^{41}0.067	0.001	0.002	0.106	0.012
12	ANKE-0	^{86}0.016	^{67}0.038	^{24}0.897	^{112}0.289	^{87}0.065	^{80}0.151	^{24}0.986		0.000	0.001	0.080	0.001
13	ANKE-1	^{87}0.016	^{66}0.038		^{111}0.284	^{88}0.065	^{81}0.151			0.000	0.001		0.001
14	AWARE-0	^{145}0.064	^{132}0.138		^{125}0.588	^{126}0.128	^{131}0.253		^{122}0.587	0.006	0.054		0.143
15	AWARE-1	^{141}0.059			^{124}0.580	^{125}0.127			^{121}0.580	0.006			0.143
16	AWARE-2	^{142}0.060				^{122}0.120				0.006			0.143
17	AWARE-3	^{116}0.033	^{112}0.090		^{122}0.503	^{101}0.085	^{100}0.204		^{118}0.505	0.004	0.003		0.014
18	AWARE-4	^{147}0.070	^{128}0.176			^{138}0.177	^{125}0.375			0.003	0.003		
19	AWARE-5	^{117}0.034	^{98}0.067	^{33}0.979	^{123}0.509	^{103}0.088	^{85}0.163	^{33}0.999	^{119}0.508	0.001	0.002	0.189	0.002
20	AWARE-6	^{149}0.072	^{131}0.128			^{140}0.178	^{135}0.283			0.001	0.002		
21	AYONIX-0	^{191}0.452	^{157}0.685		^{120}0.400	^{187}0.725	^{184}0.892		^{122}0.586	0.010	0.031		0.068
22	AYONIX-1	^{187}0.343	^{152}0.527		^{117}0.334	^{188}0.702	^{180}0.845		^{120}0.555	0.010	0.031		0.066
23	AYONIX-2	^{186}0.343	^{153}0.527			^{186}0.702	^{149}0.845			0.010	0.031		
24	CAMVI-1	^{179}0.227	^{143}0.337		^{86}0.148	^{178}0.549	^{144}0.648		^{95}0.196	0.005	0.009		0.058
25	CAMVI-2	^{180}0.129			^{81}0.130	^{184}0.402			^{90}0.157	0.005			0.058
26	CAMVI-3	^{140}0.054	^{113}0.090		^{84}0.139	^{88}0.060	^{40}0.108		^{77}0.130	0.006	0.013		0.074
27	CAMVI-4	^{137}0.049	^{107}0.077	^{16}0.640	^{136}1.000	^{79}0.056	^{58}0.100	^{30}0.994	^{134}1.000	0.000	0.000	0.000	0.000
28	CAMVI-5	^{146}0.067	^{117}0.103		^{137}1.000	^{99}0.078	^{75}0.132		^{136}1.000	0.000	0.000		0.001
29	COGENT-0	^{74}0.013	^{89}0.046		^{78}0.093	^{57}0.032	^{62}0.100		^{72}0.110	0.000	0.000		0.000
30	COGENT-1	^{73}0.013	^{91}0.048			^{51}0.032	^{61}0.100			0.000	0.000		
31	COGENT-2	^{26}0.006	^{37}0.020	^{25}0.901	^{21}0.046	^{17}0.020	^{25}0.063	^{26}0.993	^{33}0.051	0.000	0.000	0.000	0.000
32	COGENT-3	^{27}0.006	^{38}0.021		^{33}0.053	^{18}0.018	^{23}0.061		^{32}0.063	0.000	0.000		0.000
33	COGNITEC-0	^{113}0.028	^{92}0.059			^{108}0.098	^{98}0.200			0.003	0.002		
34	COGNITEC-1	^{83}0.014	^{67}0.034		^{61}0.074	^{77}0.055	^{77}0.136		^{46}0.072	0.003	0.002		0.025
35	COGNITEC-2	^{42}0.008	^{49}0.026	^{29}0.941	^{30}0.065	^{42}0.027	^{64}0.101	^{15}0.947	^{28}0.061	0.003	0.002	0.924	0.021
36	COGNITEC-3	^{45}0.009	^{48}0.025		^{29}0.051	^{44}0.028	^{59}0.100		^{13}0.049	0.004	0.002		0.012
37	DAHUA-0	^{64}0.012	^{51}0.026			^{70}0.047	^{47}0.083			0.004	0.003		
38	DAHUA-1	^{47}0.009	^{49}0.024	^{13}0.590	^{4}0.038	^{38}0.039	^{40}0.075	^{10}0.862	^{6}0.043	0.002	0.002	0.346	0.001
39	DERMALOG-0	^{161}0.131	^{139}0.218		^{63}0.075	^{157}0.364	^{136}0.528		^{69}0.104	0.003	0.002		0.020
40	DERMALOG-1	^{163}0.156			^{75}0.089	^{165}0.405			^{81}0.131	0.003			0.020
41	DERMALOG-2	^{162}0.138			^{66}0.076	^{161}0.378			^{70}0.105	0.003			0.020
42	DERMALOG-3	^{158}0.128	^{142}0.217			^{158}0.362	^{138}0.526			0.002	0.002		
43	DERMALOG-4	^{157}0.127	^{141}0.215		^{58}0.066	^{157}0.360	^{137}0.526		^{61}0.095	0.001	0.002		0.013
44	DERMALOG-5	^{89}0.017	^{66}0.037		^{57}0.066	^{64}0.045	^{56}0.096		^{38}0.066	0.001	0.002		0.013
45	DERMALOG-6	^{56}0.010	^{47}0.024	^{13}0.517	^{35}0.056	^{46}0.028	^{39}0.067	^{9}0.856	^{36}0.054	0.003	0.006	0.181	0.014
46	EVERAI-0	^{99}0.021	^{69}0.038			^{73}0.047	^{60}0.100			0.000	0.000		
47	EVERAI-1	^{20}0.006	^{35}0.020		^{129}0.928	^{27}0.023	^{36}0.074		^{127}0.927	0.000	0.000		0.000
48	EVERAI-2	^{22}0.006	^{36}0.022		^{113}0.302	^{34}0.025	^{41}0.076		^{109}0.308	0.000	0.000		0.001
49	EVERAI-3	^{15}0.005	^{28}0.019	^{4}0.154	^{5}0.038	^{17}0.018	^{21}0.060	^{6}0.535	^{11}0.044	0.000	0.000	0.032	0.001
50	EYEDEA-0	^{184}0.300	^{149}0.443		^{92}0.131	^{184}0.679	^{147}0.783		^{112}0.249	0.001	0.003		0.008
51	EYEDEA-1	^{172}0.198			^{60}0.072	^{169}0.480			^{80}0.131	0.001			0.008
52	EYEDEA-2	^{173}0.200			^{57}0.070	^{172}0.490			^{78}0.130	0.000			0.005
53	EYEDEA-3	^{151}0.082	^{129}0.148		^{46}0.064	^{150}0.267	^{136}0.404		^{56}0.091	0.001	0.003		0.008
54	GLORY-0	^{168}0.180	^{140}0.320			^{154}0.297	^{130}0.470			0.011	0.013		
55	GLORY-1	^{169}0.129	^{137}0.267		^{114}0.315	^{147}0.238	^{128}0.448		^{110}0.353	0.011	0.013		0.114
56	GORILLA-0				^{133}0.994				^{131}0.994	0.001			0.008
57	GORILLA-1	^{143}0.063	^{115}0.095		^{37}0.057	^{148}0.248	^{119}0.514		^{48}0.076	0.001	0.001		0.007
58	GORILLA-2	^{100}0.022	^{79}0.044		^{19}0.045	^{114}0.108	^{90}0.170		^{10}0.049	0.001	0.001		0.006
59	GORILLA-3	^{121}0.038	^{101}0.070		^{38}0.069	^{133}0.160	^{110}0.247		^{51}0.080	0.001	0.001		0.007
60	HBINNO-0	^{183}0.275			^{118}0.335	^{182}0.632			^{112}0.411	0.007			0.151
61	HIK-0	^{107}0.024	^{80}0.033		^{87}0.153	^{93}0.070	^{66}0.106		^{87}0.155	0.010	0.004		0.027
62	HIK-1	^{91}0.017			^{101}0.162	^{91}0.067			^{92}0.166	0.002			0.013
63	HIK-2	^{90}0.017			^{83}0.094	^{92}0.067			^{66}0.103	0.001			0.008
64	HIK-3	^{82}0.014	^{53}0.027			^{82}0.060	^{67}0.105			0.000	0.000		
65	HIK-4	^{80}0.014	^{50}0.027		^{42}0.062	^{80}0.056	^{63}0.101		^{47}0.075	0.000	0.000		0.008
66	HIK-5	^{29}0.007	^{16}0.017	^{10}0.371		^{28}0.022	^{11}0.048	^{29}0.994		0.000	0.000	0.000	0.001
67	HIK-6	^{30}0.007	^{15}0.017	^{11}0.371	^{135}1.000	^{26}0.022	^{14}0.052	^{32}0.997	^{133}1.000	0.000	0.000	0.000	0.001
68	IDEMIA-0	^{61}0.011	^{66}0.034		^{104}0.166	^{86}0.062	^{83}0.156		^{107}0.288	0.003	0.000		0.002
69	IDEMIA-1	^{65}0.012			^{99}0.157	^{50}0.031			^{97}0.205	0.003			0.002
70	IDEMIA-2	^{71}0.013			^{107}0.198	^{53}0.032			^{100}0.242	0.005			0.031
71	IDEMIA-3	^{54}0.010	^{61}0.024			^{32}0.024	^{42}0.079			0.000	0.000		
72	IDEMIA-4	^{50}0.009	^{59}0.032	^{27}0.924	^{27}0.051	^{31}0.024	^{42}0.079	^{17}0.962	^{35}0.064	0.000	0.000	0.041	0.003

Table 22: **Miss rates by dataset:** *At left, rank 1 miss rates relevant to investigations; at right, with threshold set to target FPIR = 0.01 for higher volume, low prior, uses. +For the WILD set, FPIR = 0.1 Yellow indicates most accurate algorithm. Throughout blue superscripts indicate the rank of the algorithm for that column.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

		INVESTIGATION MODE RANK ONE MISS RATE, FNIR(N, 0, 1)				IDENTIFICATION MODE HIGH T → FPIR = 0.01, FNIR(N, T, L)				FAILURE TO EXTRACT FEATURES			
		N=1.6M	N=1.6M	N=1.6M	N=1.1M	N=1.6M	N=1.6M	N=1.6M	N=1.1M	N=1.6M	N=1.6M	N=1.6M	N=1.1M
#	ALGORITHM	FRVT-18	WEBCAM	PROFILE	WILD	FRVT-18	WEBCAM	PROFILE	WILD*	FRVT-18	WEBCAM	PROFILE	WILD
73	IDEMIA-5	^{29}0.011	^{72}0.029	^{30}0.942	^{16}0.044	^{40}0.028	^{80}0.102	^{19}0.968	^{27}0.055	0.000	0.000	0.041	0.000
74	IDEMIA-6	^{30}0.012	^{108}0.072		^{31}0.052	^{38}0.028	^{89}0.161		^{39}0.067	0.000	0.000		0.000
75	IMAGUS-0	^{125}0.305	^{149}0.482		^{109}0.222	^{181}0.608	^{142}0.729		^{119}0.311	0.009	0.013		0.049
76	IMAGUS-2	^{177}0.222	^{139}0.301		^{98}0.154	^{170}0.566	^{143}0.645		^{106}0.252	0.004	0.008		0.023
77	IMAGUS-3	^{188}0.358	^{150}0.513			^{183}0.670	^{146}0.809			0.004	0.008		
78	INCODE-0	^{129}0.051	^{116}0.100			^{144}0.201	^{118}0.304			0.001	0.004		
79	INCODE-1	^{93}0.019	^{83}0.046		^{30}0.052	^{115}0.114	^{96}0.198		^{31}0.062	0.001	0.004		0.009
80	INCODE-2	^{98}0.020	^{86}0.048		^{8}0.039	^{112}0.102	^{91}0.176		^{13}0.045	0.000	0.001		0.001
81	INCODE-3	^{89}0.015	^{79}0.040		^{10}0.039	^{101}0.086	^{87}0.164		^{12}0.044	0.000	0.001		0.001
82	INNOVATRICS-0	^{127}0.042	^{106}0.076		^{105}0.188	^{138}0.165	^{112}0.258		^{105}0.245	0.002	0.008		0.093
83	INNOVATRICS-1	^{126}0.042			^{106}0.193	^{137}0.165			^{99}0.221	0.002			0.093
84	INNOVATRICS-2	^{136}0.048	^{103}0.074			^{132}0.142	^{102}0.209			0.000	0.001		
85	INNOVATRICS-3	^{113}0.029	^{88}0.055		^{58}0.071	^{128}0.134	^{99}0.203		^{52}0.081	0.000	0.001		0.007
86	INNOVATRICS-4	^{84}0.015	^{75}0.040	^{32}0.940	^{55}0.067	^{98}0.076	^{79}0.149	^{18}0.966	^{43}0.071	0.000	0.001	0.046	0.013
87	ISYSTEMS-0	^{77}0.014	^{71}0.038		^{103}0.163	^{69}0.047	^{70}0.110		^{96}0.169	0.003	0.013		0.065
88	ISYSTEMS-1	^{76}0.014			^{102}0.162	^{67}0.047			^{95}0.169	0.003			0.065
89	ISYSTEMS-2	^{54}0.009	^{50}0.026		^{24}0.049	^{56}0.035	^{45}0.080		^{22}0.051	0.002	0.002		0.009
90	ISYSTEMS-3	^{37}0.007	^{42}0.023	^{15}0.718	^{19}0.043	^{40}0.027	^{31}0.068	^{31}0.995	^{10}0.044	0.002	0.002	0.142	0.003
91	LOOKMAN-3	^{82}0.011	^{70}0.038		^{181}1.000	^{41}0.027	^{49}0.082			0.000	0.000		0.000
92	LOOKMAN-4	^{66}0.012	^{72}0.039	^{32}0.978	^{188}1.000	^{39}0.027	^{39}0.075	^{22}0.978		0.000	0.000	0.000	0.000
93	MEGVII-0	^{51}0.009	^{18}0.017		^{41}0.061	^{51}0.058	^{30}0.067		^{60}0.094	0.000	0.000		0.005
94	MEGVII-1	^{78}0.014	^{19}0.017			^{57}0.039	^{22}0.061			0.002	0.000		
95	MEGVII-2	^{79}0.014	^{20}0.017	^{7}0.275		^{59}0.039	^{20}0.059	^{8}0.698		0.002	0.000	0.033	
96	MICROFOCUS-0	^{138}0.597	^{161}0.782		^{115}0.316	^{152}0.867	^{157}0.950		^{115}0.434	0.005	0.030		0.065
97	MICROFOCUS-1	^{136}0.597			^{116}0.316	^{153}0.867			^{116}0.434	0.005			0.065
98	MICROFOCUS-2	^{137}0.627			^{119}0.342	^{154}0.870			^{117}0.447	0.005			0.065
99	MICROFOCUS-3	^{134}0.595	^{160}0.781		^{110}0.279	^{151}0.866	^{156}0.948		^{113}0.412	0.001	0.005		0.014
100	MICROFOCUS-4	^{133}0.577	^{159}0.758			^{159}0.999	^{155}0.940			0.001	0.005		
101	MICROFOCUS-5	^{129}0.426	^{136}0.601		^{100}0.158	^{149}0.736	^{153}0.865		^{116}0.261	0.001	0.005		0.011
102	MICROFOCUS-6	^{130}0.428	^{135}0.583		^{95}0.146	^{175}0.963	^{151}0.858		^{102}0.246	0.001	0.005		0.011
103	MICROSOFT-0	^{23}0.006	^{31}0.021		^{39}0.065	^{22}0.022	^{33}0.071		^{37}0.065	0.000	0.001		0.019
104	MICROSOFT-1	^{21}0.006			^{44}0.062	^{24}0.022			^{29}0.061	0.000			0.019
105	MICROSOFT-2	^{25}0.006			^{45}0.063	^{26}0.026			^{34}0.063	0.000			0.019
106	MICROSOFT-3	^{4}0.003	^{8}0.012			^{16}0.014	^{18}0.056			0.000	0.001		
107	MICROSOFT-4	^{2}0.003	^{7}0.012		^{9}0.039	^{15}0.013	^{15}0.053		^{9}0.043	0.000	0.001		0.004
108	MICROSOFT-5	^{5}0.003	^{5}0.011	^{1}0.087	^{2}0.033	^{12}0.012	^{8}0.041	^{2}0.183	^{4}0.041	0.000	0.001	0.049	0.000
109	MICROSOFT-6	^{8}0.003	^{4}0.011	^{2}0.089		^{5}0.008	^{6}0.024	^{1}0.148		0.000	0.001	0.049	
110	NEC-0	^{34}0.020	^{77}0.041		^{134}0.999	^{74}0.049	^{52}0.093		^{132}0.999	0.001	0.002		0.064
111	NEC-1	^{106}0.024	^{89}0.056			^{87}0.063	^{76}0.133			0.005	0.003		
112	NEC-2	^{1}0.003	^{2}0.009		^{80}0.093	^{1}0.004	^{2}0.003		^{71}0.107	0.000	0.001		0.025
113	NEC-3	^{3}0.003	^{3}0.010	^{6}0.272	^{74}0.088	^{2}0.004	^{1}0.013	^{3}0.479	^{58}0.092	0.000	0.001	0.041	0.025
114	NEUROTECHNOLOGY-0	^{138}0.050	^{113}0.104		^{137}1.000	^{143}0.196	^{120}0.317		^{135}1.000	0.004	0.022		0.091
115	NEUROTECHNOLOGY-1	^{135}0.047			^{130}0.954	^{142}0.195			^{128}0.953	0.001			0.028
116	NEUROTECHNOLOGY-2	^{134}0.047			^{131}0.983	^{140}0.195			^{129}0.983	0.001			0.029
117	NEUROTECHNOLOGY-3	^{109}0.025	^{76}0.042			^{113}0.101	^{86}0.164			0.000	0.001		
118	NEUROTECHNOLOGY-4	^{40}0.008	^{35}0.020		^{76}0.090	^{46}0.030	^{34}0.073		^{74}0.122	0.000	0.001		0.007
119	NEUROTECHNOLOGY-5	^{31}0.007	^{46}0.024	^{22}0.854	^{121}0.408	^{35}0.025	^{37}0.074	^{23}0.982	^{114}0.415	0.000	0.000	0.030	0.000
120	NEUROTECHNOLOGY-6	^{98}0.020	^{80}0.045		^{25}0.050	^{123}0.124	^{101}0.206		^{36}0.065	0.000	0.000		0.001
121	NEWLAND-2	^{150}0.081	^{119}0.117			^{159}0.296	^{122}0.335			0.007	0.012		
122	NOBLIS-1	^{161}0.251	^{151}0.522		^{127}0.734	^{177}0.992	^{159}1.000		^{124}0.744	0.000	0.000		0.000
123	NOBLIS-2	^{159}0.182	^{146}0.392	^{31}0.971		^{173}0.490	^{166}1.000	^{34}1.000		0.000	0.000	0.000	
124	NTECHLAB-0	^{63}0.012	^{58}0.031		^{12}0.041	^{72}0.047	^{68}0.105		^{7}0.045	0.000	0.001		0.005
125	NTECHLAB-1	^{81}0.014			^{30}0.045	^{78}0.056			^{21}0.049	0.000			0.005
126	NTECHLAB-3	^{41}0.008	^{39}0.023			^{47}0.030	^{37}0.075			0.000	0.000		
127	NTECHLAB-4	^{33}0.007	^{32}0.019		^{14}0.043	^{29}0.024	^{26}0.065		^{18}0.048	0.000	0.000		0.003
128	NTECHLAB-5	^{28}0.006	^{21}0.018		^{7}0.038	^{30}0.024	^{28}0.063		^{6}0.042	0.000	0.000		0.000
129	NTECHLAB-6	^{24}0.006	^{17}0.017	^{5}0.208	^{6}0.038	^{20}0.021	^{19}0.059	^{4}0.443	^{5}0.042	0.000	0.000	0.040	0.000
130	QUANTASOFT-1	^{176}0.220	^{158}0.727		^{126}0.620	^{178}0.494			^{125}0.760	0.000	0.000		0.000
131	RANKONE-0	^{133}0.045	^{120}0.117		^{89}0.114	^{127}0.129	^{116}0.291		^{72}0.161	0.000	0.000		0.000
132	RANKONE-1	^{108}0.025			^{88}0.077	^{102}0.087			^{67}0.102	0.000			0.000
133	RANKONE-2	^{102}0.022	^{102}0.071			^{97}0.073	^{95}0.190			0.000	0.000		
134	RANKONE-3	^{101}0.022	^{97}0.068		^{69}0.078	^{96}0.073	^{93}0.187		^{62}0.095	0.000	0.000		0.000
135	RANKONE-4	^{132}0.044	^{123}0.141		^{82}0.094	^{124}0.126	^{121}0.324		^{75}0.126	0.000	0.000		0.000
136	RANKONE-5	^{68}0.012	^{56}0.041	^{38}0.981	^{39}0.061	^{90}0.086	^{72}0.119	^{28}0.994	^{47}0.068	0.000	0.000	0.489	0.000
137	REALNETWORKS-0	^{131}0.042	^{110}0.078		^{60}0.076	^{131}0.140	^{104}0.209		^{58}0.084	0.001	0.000		0.004
138	REALNETWORKS-1	^{130}0.043	^{105}0.078			^{131}0.140	^{105}0.209			0.001	0.000		
139	REALNETWORKS-2	^{125}0.042	^{108}0.078		^{132}0.992	^{129}0.139	^{106}0.209		^{130}0.992	0.001	0.000		0.000
140	REMARKAI-0	^{41}0.011	^{37}0.030			^{88}0.062	^{74}0.123			0.000	0.001		
141	REMARKAI-2	^{38}0.010	^{35}0.029	^{19}0.708	^{23}0.046	^{84}0.061	^{73}0.122	^{16}0.958	^{25}0.052	0.000	0.001	0.017	0.000
142	SENSETIME-0	^{16}0.005	^{13}0.016	^{12}0.446		^{10}0.012	^{6}0.040	^{30}0.971		0.004	0.000	0.042	0.000
143	SENSETIME-1	^{17}0.005	^{12}0.016		^{3}0.038	^{11}0.012	^{10}0.041		^{1}0.796	0.004	0.000		0.000
144	SHAMAN-0	^{186}0.171	^{136}0.262		^{30}0.115	^{160}0.370	^{134}0.507		^{98}0.146	0.020	0.011		0.043

Table 23: **Miss rates by dataset:** *At left, rank 1 miss rates relevant to investigations; at right, with threshold set to target FPIR = 0.01 for higher volume, low prior, uses. *For the WILD set, FPIR = 0.1 Yellow indicates most accurate algorithm. Throughout blue superscripts indicate the rank of the algorithm for that column.*

		INVESTIGATION MODE				IDENTIFICATION MODE				FAILURE TO EXTRACT FEATURES			
		RANK ONE MISS RATE, FNIR(N, 0, 1)				HIGHT → FPIR = 0.01, FNIR(N, T, L)							
		N=1.6M	N=1.6M	N=1.6M	N=1.1M	N=1.6M	N=1.6M	N=1.6M	N=1.1M	N=1.6M	N=1.6M	N=1.6M	N=1.1M
#	ALGORITHM	FRVT-18	WEBCAM	PROFILE	WILD	FRVT-18	WEBCAM	PROFILE	WILD⁺	FRVT-18	WEBCAM	PROFILE	WILD
145	SHAMAN-1	^{166}0.172			^{88}0.113	^{166}0.406			^{88}0.153	0.020			0.043
146	SHAMAN-2	^{182}0.262			^{90}0.132	^{180}0.582			^{96}0.201	0.020			0.043
147	SHAMAN-3	^{158}0.127	^{137}0.172		^{88}0.109	^{156}0.348	^{131}0.472		^{82}0.132	0.020	0.011		0.043
148	SHAMAN-4	^{178}0.224	^{139}0.319			^{171}0.490	^{142}0.639			0.020	0.011		
149	SHAMAN-6	^{125}0.042	^{71}0.058	^{26}0.910		^{105}0.095	^{83}0.168	^{14}0.935		0.020	0.011	0.869	
150	SHAMAN-7	^{123}0.042	^{90}0.057		^{70}0.078	^{104}0.094	^{85}0.169		^{80}0.079	0.020	0.010		0.029
151	SIAT-0	^{56}0.010	^{32}0.021		^{71}0.078	^{63}0.047	^{27}0.064		^{134}0.250	0.000	0.000		0.008
152	SIAT-1	^{10}0.004	^{142}0.333		^{11}0.040	^{6}0.009	^{133}0.348		^{3}0.041	0.000	0.000		0.003
153	SIAT-2	^{11}0.004	^{148}0.446			^{7}0.009	^{120}0.460			0.000	0.000		
154	SMILART-0	^{170}0.198	^{140}0.325		^{195}1.000	^{170}0.486			^{136}1.000	0.008			0.121
155	SMILART-1	^{175}0.219			^{154}1.000	^{177}0.505			^{163}1.000	0.021			0.006
156	SMILART-2	^{171}0.196			^{140}1.000	^{174}0.492			^{144}1.000	0.000			0.048
157	SMILART-4	^{196}0.965	^{162}0.974		^{136}0.834	^{196}0.965	^{150}0.973		^{126}0.833	0.011	0.013		0.089
158	SMILART-5									0.011	0.013		
159	SYNESIS-0	^{164}0.162	^{145}0.361			^{162}0.378	^{141}0.598			0.002	0.009		0.081
160	SYNESIS-3	^{167}0.172	^{134}0.235			^{168}0.444	^{135}0.524			0.006	0.015		0.042
161	TEVIAN-0	^{104}0.022	^{37}0.066		^{34}0.054	^{117}0.114	^{106}0.227		^{44}0.072	0.002	0.005		0.007
162	TEVIAN-1	^{105}0.022			^{43}0.062	^{118}0.114			^{45}0.078	0.002			0.007
163	TEVIAN-2	^{106}0.022			^{78}0.093	^{116}0.114			^{74}0.118	0.002			0.008
164	TEVIAN-3	^{86}0.017	^{86}0.052			^{109}0.098	^{97}0.198			0.001	0.002		
165	TEVIAN-4	^{75}0.013	^{63}0.038		^{36}0.050	^{80}0.066	^{71}0.115		^{43}0.063	0.001	0.002		0.005
166	TEVIAN-5	^{46}0.009	^{54}0.028	^{8}0.329		^{71}0.047	^{45}0.089	^{7}0.661		0.001	0.002	0.116	
167	TIGER-0	^{144}0.064	^{115}0.095		^{186}1.000	^{149}0.263	^{124}0.366		^{165}1.000	0.000	0.000		0.005
168	TIGER-1		^{144}0.351				^{132}0.487			0.000	0.000		
169	TIGER-2	^{36}0.008	^{41}0.023	^{7}0.355		^{61}0.042	^{55}0.095	^{17}0.927		0.000	0.000	0.056	
170	TIGER-3	^{36}0.008	^{40}0.023			^{62}0.042	^{54}0.095			0.000	0.000		
171	TONGYITRANS-0	^{53}0.010	^{38}0.022			^{60}0.041	^{32}0.069			0.003	0.001		
172	TONGYITRANS-1	^{52}0.010	^{37}0.022		^{87}0.112	^{54}0.036	^{28}0.062		^{83}0.134	0.003	0.001		0.009
173	TOSHIBA-0	^{32}0.007	^{34}0.022	^{17}0.689		^{47}0.029	^{35}0.074	^{21}0.971		0.000	0.000	0.070	0.002
174	TOSHIBA-1	^{34}0.007	^{36}0.022			^{31}0.021	^{16}0.054			0.000	0.000		
175	VD-0	^{192}0.475	^{154}0.551		^{106}0.217	^{190}0.828	^{153}0.871		^{111}0.362	0.011	0.018		0.026
176	VD-1	^{135}0.030	^{67}0.053			^{121}0.118	^{94}0.188			0.005	0.001		0.017
177	VIGILANTSOLUTIONS-0	^{154}0.125	^{130}0.212		^{65}0.076	^{153}0.394	^{140}0.557		^{97}0.152	0.000	0.001		0.003
178	VIGILANTSOLUTIONS-1	^{174}0.204			^{86}0.108	^{170}0.502			^{98}0.209	0.000			0.003
179	VIGILANTSOLUTIONS-2	^{180}0.239			^{47}0.064	^{188}0.731			^{76}0.129	0.000			0.003
180	VIGILANTSOLUTIONS-3	^{145}0.072	^{128}0.151		^{51}0.065	^{151}0.283	^{137}0.526		^{79}0.131	0.000	0.001		0.003
181	VIGILANTSOLUTIONS-4	^{158}0.127	^{138}0.244			^{157}0.424	^{145}0.709			0.000	0.001		
182	VIGILANTSOLUTIONS-5	^{67}0.012				^{53}0.045				0.000	0.001		
183	VIGILANTSOLUTIONS-6	^{70}0.013				^{65}0.046				0.000	0.001		
184	VISIONLABS-3	^{48}0.009	^{58}0.030		^{28}0.051	^{37}0.026	^{50}0.091		^{15}0.046	0.002	0.003		0.014
185	VISIONLABS-4	^{13}0.004	^{25}0.020			^{38}0.026	^{57}0.097			0.001	0.001		
186	VISIONLABS-5	^{12}0.004	^{23}0.019		^{18}0.043	^{25}0.022	^{46}0.087		^{16}0.046	0.001	0.001		0.006
187	VISIONLABS-6	^{7}0.003	^{15}0.015			^{14}0.012	^{13}0.051			0.001	0.001		
188	VISIONLABS-7	^{5}0.003	^{10}0.015	^{3}0.130	^{1}0.033	^{13}0.012	^{12}0.051	^{3}0.322	^{2}0.035	0.001	0.001	0.051	0.001
189	VOCORD-0	^{129}0.040	^{100}0.068			^{120}0.116	^{92}0.181			0.015	0.025		0.019
190	VOCORD-1	^{122}0.040				^{119}0.116				0.015			0.018
191	VOCORD-2	^{120}0.038				^{113}0.107				0.015			0.015
192	VOCORD-3	^{40}0.008	^{45}0.024		^{36}0.057	^{75}0.050	^{52}0.093		^{31}0.061	0.001	0.011		0.006
193	VOCORD-4	^{57}0.010	^{31}0.021			^{76}0.054	^{51}0.093			0.000	0.000		
194	VOCORD-5	^{45}0.009	^{43}0.023	^{20}0.739	^{17}0.044	^{66}0.046	^{44}0.080	^{13}0.929	^{14}0.045	0.001	0.009	0.554	0.003
195	VOCORD-6	^{205}1.000	^{201}1.000			^{203}1.000	^{201}1.000			0.001	0.009		
196	YISHENG-0	^{131}0.027	^{93}0.060		^{54}0.067	^{146}0.209	^{114}0.275		^{96}0.100	0.002	0.005		0.014
197	YISHENG-1	^{114}0.029	^{84}0.060		^{40}0.061	^{145}0.208	^{113}0.269		^{54}0.087	0.002	0.005		0.014
198	YITU-0	^{38}0.007	^{29}0.020		^{72}0.086	^{33}0.025	^{17}0.054		^{55}0.094	0.003	0.001		0.026
199	YITU-1	^{35}0.007			^{72}0.086	^{28}0.023			^{57}0.092	0.003			0.026
200	YITU-2	^{14}0.004	^{4}0.010		^{22}0.046	^{8}0.011	^{6}0.028		^{24}0.051	0.000	0.000		0.000
201	YITU-3	^{15}0.005	^{16}0.016			^{9}0.011	^{7}0.033			0.003	0.001		
202	YITU-4	^{9}0.004	^{1}0.008	^{21}0.831	^{18}0.044	^{3}0.007	^{3}0.017	^{11}0.875	^{17}0.047	0.000	0.000	0.000	0.006
203	YITU-5	^{16}0.005	^{9}0.014			^{4}0.007	^{4}0.023			0.003	0.001		

Table 24: **Miss rates by dataset:** *At left, rank 1 miss rates relevant to investigations; at right, with threshold set to target FPIR = 0.01 for higher volume, low prior, uses. ⁺For the WILD set, FPIR = 0.1 Yellow indicates most accurate algorithm. Throughout blue superscripts indicate the rank of the algorithm for that column.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

MISSES OUTSIDE RANK R		MUGSHOT SEARCHES, N = 1.6M IDENTITIES							
FNIR(N, T, R)		INVESTIGATION MODE, T = 0 PROPORTION MATED SEARCHES				IDENTIFICATION MODE, T > 0 FOR FPIR = 0.001 PROPORTION MATED SEARCHES			
GALLERY		WITHOUT THE MATE AT RANK 1		WITH NO MATE AT RANK 1	WITH K-TH MATE NOT IN TOP K	WITH THE MATE BELOW THRESHOLD		WITHOUT ANY MATE ABOVE THRESH	WITHOUT ALL MATES ABOVE THRESH
		RECENT	CONSOLIDATED	UNCONSOLIDATED		RECENT	CONSOLIDATED	UNCONSOLIDATED	
1	3DIVI-5	[49] 0.0202	[43] 0.0133	[46] 0.0133	[48] 0.0449	[48] 0.1691	[36] 0.1339	[48] 0.1339	[49] 0.3186
2	3DIVI-6	[52] 0.0265	[46] 0.0186	[51] 0.0172	[47] 0.0410	[51] 0.1705	[47] 0.1345	[49] 0.1350	[48] 0.3160
3	ALCHERA-2	[64] 0.0973	[57] 0.0914	[63] 0.0734	[63] 0.1876	[64] 0.4899	[67] 0.3756	[63] 0.4418	[63] 0.6820
4	ANKE-0	[45] 0.0158	[39] 0.0100	[43] 0.0100	[45] 0.0338	[41] 0.1199	[38] 0.0989	[43] 0.0989	[42] 0.2558
5	ANKE-1	[46] 0.0158	[40] 0.0101	[44] 0.0101	[44] 0.0337	[42] 0.1218	[40] 0.1001	[44] 0.1001	[43] 0.2581
6	AWARE-5	[54] 0.0337	[47] 0.0208	[53] 0.0230	[57] 0.0740	[60] 0.3729	[59] 0.2984	[60] 0.3777	[61] 0.6534
7	AWARE-6	[62] 0.0722	[56] 0.0538	[60] 0.0538	[61] 0.1551	[58] 0.2779	[53] 0.2419	[52] 0.2465	[59] 0.5140
8	AYONIX-1	[70] 0.3432	[62] 0.3364	[67] 0.2841	[68] 0.4764	[68] 0.8247	[61] 0.8533	[66] 0.7935	[66] 0.9037
9	AYONIX-2	[69] 0.3432	[61] 0.2606	[68] 0.2841	[67] 0.4763	[67] 0.8246	[59] 0.8038	[65] 0.7933	[65] 0.9036
10	CAMVI-4	[50] 0.0490	[54] 0.0326	[59] 0.0469	[49] 0.0475	[31] 0.0741	[27] 0.0505	[34] 0.0661	[16] 0.1105
11	CAMVI-5	[61] 0.0673	[55] 0.0458	[62] 0.0633	[56] 0.0638	[40] 0.1020	[36] 0.0727	[40] 0.0922	[27] 0.1513
12	COGENT-2	[14] 0.0062	[12] 0.0027	[12] 0.0027	[12] 0.0086	[19] 0.0475	[12] 0.0299	[19] 0.0391	[20] 0.1275
13	COGENT-3	[15] 0.0064	[18] 0.0037	[13] 0.0029	[13] 0.0091	[21] 0.0515	[17] 0.0341	[27] 0.0450	[28] 0.1448
14	COGNITEC-2	[25] 0.0083	[22] 0.0044	[22] 0.0043	[24] 0.0145	[25] 0.0560	[22] 0.0401	[22] 0.0400	[25] 0.1342
15	COGNITEC-3	[26] 0.0088	[24] 0.0048	[28] 0.0048	[25] 0.0148	[24] 0.0555	[21] 0.0397	[21] 0.0397	[24] 0.1322
16	DAHUA-0	[38] 0.0115	[32] 0.0070	[36] 0.0072	[34] 0.0204	[37] 0.0891	[32] 0.0624	[35] 0.0691	[38] 0.1967
17	DAHUA-1	[27] 0.0089	[25] 0.0049	[27] 0.0052	[27] 0.0173	[33] 0.0755	[28] 0.0521	[30] 0.0577	[33] 0.1739
18	DERMALOG-5	[47] 0.0171	[41] 0.0113	[48] 0.0139	[41] 0.0254	[39] 0.0909	[33] 0.0649	[38] 0.0767	[39] 0.2072
19	DERMALOG-6	[30] 0.0102	[28] 0.0060	[30] 0.0061	[19] 0.0119	[23] 0.0542	[20] 0.0383	[24] 0.0416	[21] 0.1280
20	EVERAI-2	[12] 0.0058	[13] 0.0029	[16] 0.0032	[15] 0.0099	[22] 0.0526	[19] 0.0370	[23] 0.0410	[22] 0.1312
21	EVERAI-3	[8] 0.0047	[11] 0.0023	[11] 0.0024	[11] 0.0073	[11] 0.0277	[11] 0.0256	[11] 0.0285	[11] 0.0978
22	GORILLA-2	[51] 0.0220	[44] 0.0137	[50] 0.0153	[58] 0.0570	[53] 0.1902	[49] 0.1379	[57] 0.1537	[57] 0.3589
23	GORILLA-3	[55] 0.0384	[49] 0.0245	[55] 0.0283	[60] 0.1032	[59] 0.3260	[58] 0.2730	[59] 0.3043	[60] 0.5786
24	HIK-5	[17] 0.0067		[20] 0.0037	[23] 0.0140	[17] 0.0467		[18] 0.0364	[18] 0.1228
25	HIK-6	[18] 0.0067	[16] 0.0034	[17] 0.0037	[22] 0.0140	[20] 0.0500	[16] 0.0324	[20] 0.0392	[22] 0.1310
26	IDEMIA-5	[32] 0.0107	[29] 0.0062	[32] 0.0064	[33] 0.0192	[16] 0.0465	[15] 0.0319	[17] 0.0348	[17] 0.1125
27	IDEMIA-6	[37] 0.0122	[33] 0.0071	[33] 0.0076	[32] 0.0188	[14] 0.0458	[14] 0.0316	[14] 0.0342	[13] 0.1032
28	INCODE-2	[50] 0.0203	[42] 0.0120	[47] 0.0137	[50] 0.0480	[52] 0.1861	[48] 0.1360	[51] 0.1507	[51] 0.3600
29	INCODE-3	[44] 0.0153	[36] 0.0088	[45] 0.0102	[46] 0.0368	[50] 0.1703	[45] 0.1227	[50] 0.1388	[50] 0.3290
30	INNOVATRICS-4	[43] 0.0149	[35] 0.0081	[40] 0.0081	[43] 0.0293	[46] 0.1340	[37] 0.0928	[41] 0.0927	[41] 0.2479
31	ISYSTEMS-3	[22] 0.0075	[20] 0.0040	[22] 0.0041	[16] 0.0106	[29] 0.0620	[23] 0.0402	[28] 0.0500	[30] 0.1519
32	LOOKMAN-3	[34] 0.0114	[27] 0.0089	[34] 0.0067	[17] 0.0109	[15] 0.0463	[25] 0.0425	[13] 0.0338	[12] 0.1015
33	LOOKMAN-4	[36] 0.0117	[28] 0.0091	[35] 0.0072	[21] 0.0134	[18] 0.0472	[24] 0.0417	[15] 0.0346	[14] 0.1036
34	MEGVII-1	[41] 0.0137		[41] 0.0096	[36] 0.0231	[32] 0.0746		[31] 0.0577	[32] 0.1688
35	MEGVII-2	[42] 0.0137		[42] 0.0097	[38] 0.0236	[34] 0.0796		[33] 0.0623	[34] 0.1810
36	MICROFOCUS-5	[71] 0.4257	[63] 0.3701	[69] 0.3701	[69] 0.5522	[69] 0.8361	[63] 0.9835	[67] 0.8139	[67] 0.9189
37	MICROFOCUS-6	[72] 0.4283	[64] 0.3732	[70] 0.3732	[70] 0.5566	[71] 0.9780	[60] 0.8195	[68] 0.8195	[68] 0.9215
38	MICROSOFT-5	[3] 0.0033	[3] 0.0013	[6] 0.0015	[10] 0.0062	[5] 0.0279	[7] 0.0171	[7] 0.0193	[9] 0.0755
39	MICROSOFT-6	[6] 0.0033	[5] 0.0014	[7] 0.0015	[9] 0.0060	[6] 0.0141	[5] 0.0080	[10] 0.0213	[10] 0.0772
40	NEC-2	[1] 0.0028	[2] 0.0011	[1] 0.0008	[1] 0.0019	[2] 0.0047	[2] 0.0024	[1] 0.0021	[2] 0.0086
41	NEC-3	[2] 0.0032	[4] 0.0013	[2] 0.0010	[2] 0.0019	[1] 0.0044	[1] 0.0021	[2] 0.0022	[1] 0.0080
42	NEUROTECHNOLOGY-5	[19] 0.0068	[21] 0.0042	[14] 0.0032	[14] 0.0094	[26] 0.0564	[29] 0.0527	[25] 0.0438	[26] 0.1364
43	NEUROTECHNOLOGY-6	[48] 0.0201	[45] 0.0153	[42] 0.0142	[52] 0.0534	[57] 0.2555	[51] 0.2695	[57] 0.2125	[57] 0.4458
44	NEWLAND-2	[63] 0.0811		[61] 0.0599	[62] 0.1562	[63] 0.4405		[61] 0.3790	[60] 0.6252
45	NOBLIS-1	[68] 0.2512	[60] 0.2049	[65] 0.2032	[66] 0.3681	[73] 0.9996	[66] 0.9998	[72] 0.9994	[72] 0.9997
46	NOBLIS-2	[66] 0.1816	[59] 0.1565	[66] 0.2517	[65] 0.3944	[72] 0.9974	[65] 0.9959	[71] 0.9967	[71] 0.9987
47	NTECHLAB-5	[16] 0.0064	[19] 0.0039	[21] 0.0039	[31] 0.0179	[13] 0.0448	[18] 0.0347	[16] 0.0347	[19] 0.1235
48	NTECHLAB-6	[13] 0.0059	[14] 0.0034	[17] 0.0034	[26] 0.0154	[12] 0.0391	[13] 0.0301	[12] 0.0301	[15] 0.1088
49	QUANTASOFT-1	[67] 0.2198	[66] 0.9857	[71] 0.9426	[71] 0.9502	[66] 0.6399	[64] 0.9915	[69] 0.9640	[70] 0.9801
50	RANKONE-4	[59] 0.0441	[52] 0.0318	[58] 0.0318	[59] 0.0945	[54] 0.1951	[50] 0.1545	[53] 0.1545	[53] 0.3590
51	RANKONE-5	[39] 0.0120	[34] 0.0072	[37] 0.0072	[39] 0.0237	[35] 0.0617	[30] 0.0447	[36] 0.0447	[37] 0.1404
52	REALNETWORKS-2	[56] 0.0418	[53] 0.0320	[54] 0.0268	[58] 0.0903	[56] 0.2341	[52] 0.2049	[56] 0.1775	[58] 0.3949
53	REMARKAI-0	[33] 0.0109	[31] 0.0065	[38] 0.0065	[40] 0.0238	[44] 0.1201	[41] 0.1020	[45] 0.1020	[47] 0.2671
54	REMARKAI-2	[31] 0.0105	[30] 0.0062	[31] 0.0062	[37] 0.0235	[43] 0.1264	[39] 0.0991	[42] 0.0991	[44] 0.2615
55	SENSETIME-0	[9] 0.0048	[9] 0.0018	[5] 0.0018	[4] 0.0037	[8] 0.0234	[6] 0.0165	[5] 0.0168	[5] 0.0603
56	SENSETIME-1	[10] 0.0048	[8] 0.0018	[8] 0.0019	[7] 0.0041	[7] 0.0245	[8] 0.0175	[6] 0.0177	[6] 0.0628
57	SHAMAN-6	[58] 0.0424	[51] 0.0312	[57] 0.0312	[53] 0.0542	[45] 0.1432	[43] 0.1109	[46] 0.1109	[46] 0.2629
58	SHAMAN-7	[57] 0.0422	[50] 0.0310	[56] 0.0310	[51] 0.0529	[47] 0.1426	[44] 0.1112	[47] 0.1112	[45] 0.2624
59	SMILART-4	[73] 0.9649	[65] 0.9531	[72] 0.9722	[72] 0.9738	[70] 0.9683	[62] 0.9569	[70] 0.9740	[69] 0.9781
60	SYNESIS-3	[65] 0.1721	[58] 0.1350	[64] 0.1350	[64] 0.2571	[65] 0.5832	[58] 0.5296	[64] 0.5295	[64] 0.7459
61	TEVIAN-5	[28] 0.0092	[26] 0.0053	[29] 0.0058	[29] 0.0213	[36] 0.0898	[34] 0.0667	[39] 0.0770	[40] 0.2079
62	TIGER-2	[24] 0.0075	[23] 0.0044	[25] 0.0044	[29] 0.0177	[36] 0.0888	[35] 0.0698	[36] 0.0698	[38] 0.2016
63	TIGER-3	[23] 0.0075		[24] 0.0044	[28] 0.0177	[35] 0.0888		[37] 0.0698	[37] 0.2015
64	TOSHIBA-0	[20] 0.0068	[14] 0.0033	[16] 0.0033	[18] 0.0110	[30] 0.0648	[31] 0.0529	[29] 0.0529	[31] 0.1599
65	TOSHIBA-1	[21] 0.0071	[17] 0.0035	[15] 0.0035	[20] 0.0120	[28] 0.0618	[31] 0.0596	[31] 0.0595	[35] 0.1819
66	VD-1	[53] 0.0302	[48] 0.0221	[52] 0.0221	[54] 0.0560	[55] 0.2036	[51] 0.1654	[54] 0.1658	[54] 0.3667
67	VIGILANTSOLUTIONS-5	[37] 0.0118				[62] 0.4327			
68	VIGILANTSOLUTIONS-6	[40] 0.0125		[39] 0.0077	[42] 0.0258	[61] 0.4260		[62] 0.4165	[62] 0.6577
69	VISIONLABS-6	[3] 0.0033	[7] 0.0015	[5] 0.0015	[6] 0.0040	[10] 0.0289	[10] 0.0185	[9] 0.0201	[8] 0.0737
70	VISIONLABS-7	[4] 0.0033	[6] 0.0014	[4] 0.0014	[5] 0.0039	[9] 0.0289	[9] 0.0185	[8] 0.0201	[7] 0.0737
71	VOCORD-5	[29] 0.0092	[37] 0.0057	[28] 0.0054	[31] 0.0182	[49] 0.1697	[42] 0.1076	[55] 0.1717	[56] 0.3775
72	YITU-4	[7] 0.0037	[1] 0.0011	[3] 0.0012	[3] 0.0033	[3] 0.0123	[3] 0.0074	[3] 0.0080	[3] 0.0337
73	YITU-5	[11] 0.0048	[10] 0.0020	[10] 0.0020	[8] 0.0041	[4] 0.0128	[4] 0.0076	[4] 0.0088	[4] 0.0350

Table 25: **Comparing enrollment styles for the FRVT 2018 mugshot sets.** *Consolidated refers to enrollment of all lifetime images in one template Unconsolidated refers to enrollment of those images separately under different identifiers. Columns 3 - 6 values are FNIR at rank 1 and with* T = 0*. Columns 7 - 10 values are high threshold FNIR. Throughout, blue superscripts indicate the rank of the algorithm for that column, and the best three values are highlighted in yellow and green.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

2019/09/11 17:24:52

FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

Figure 19: **[Mugshot Dataset] Error rate reductions in 2018.** *For each* FRVT*2018 participant, the plot shows accuracy gains between Phase 1 (Feb 2018), Phase 2 (Jun 2018) and Phase 3 (Nov 2018) according to two metrics: rank one miss rate, FNIR(N, 1, 0), and high threshold, FNIR(N, L, T), with T set to achieve FPIR = 0.003. The text "Red=" gives the best reduction multiplier for the given metric on the recent enrollment strategy - a smaller value is better.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

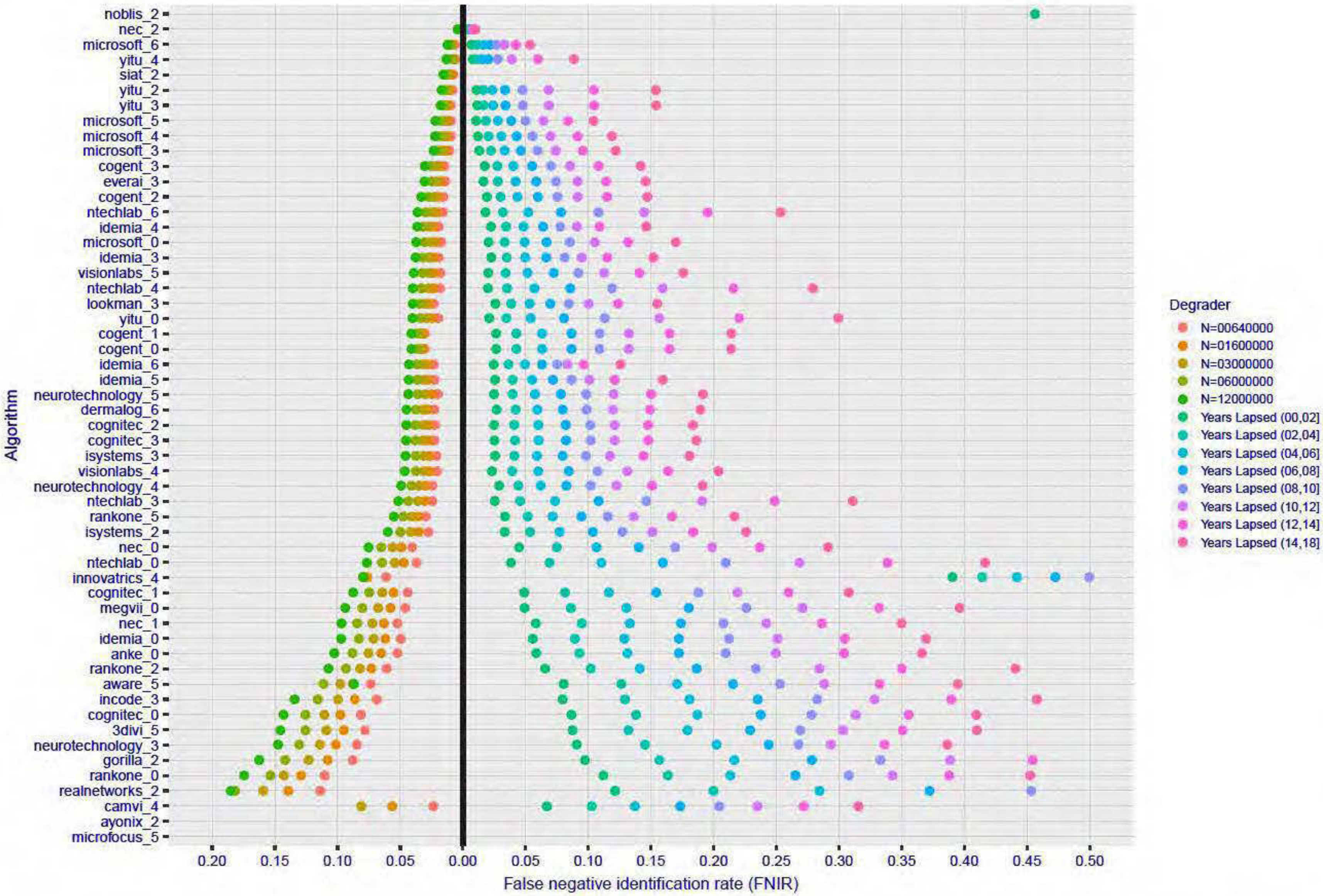


Figure 20: **[FRVT-2018 Mugshot Ageing Dataset] Contrast of ageing and population size dependency.** *The Figure shows, at left, the dependence FNIR(N) for the FRVT-2018, as tabulated in Table 12. At right, is FNIR($N = 3\,000\,000, \Delta T$) from Figure 62. Ageing miss rates are computed over all searches binned by number of years between search and initial enrollment. In all cases, FPIR = 0.01.*

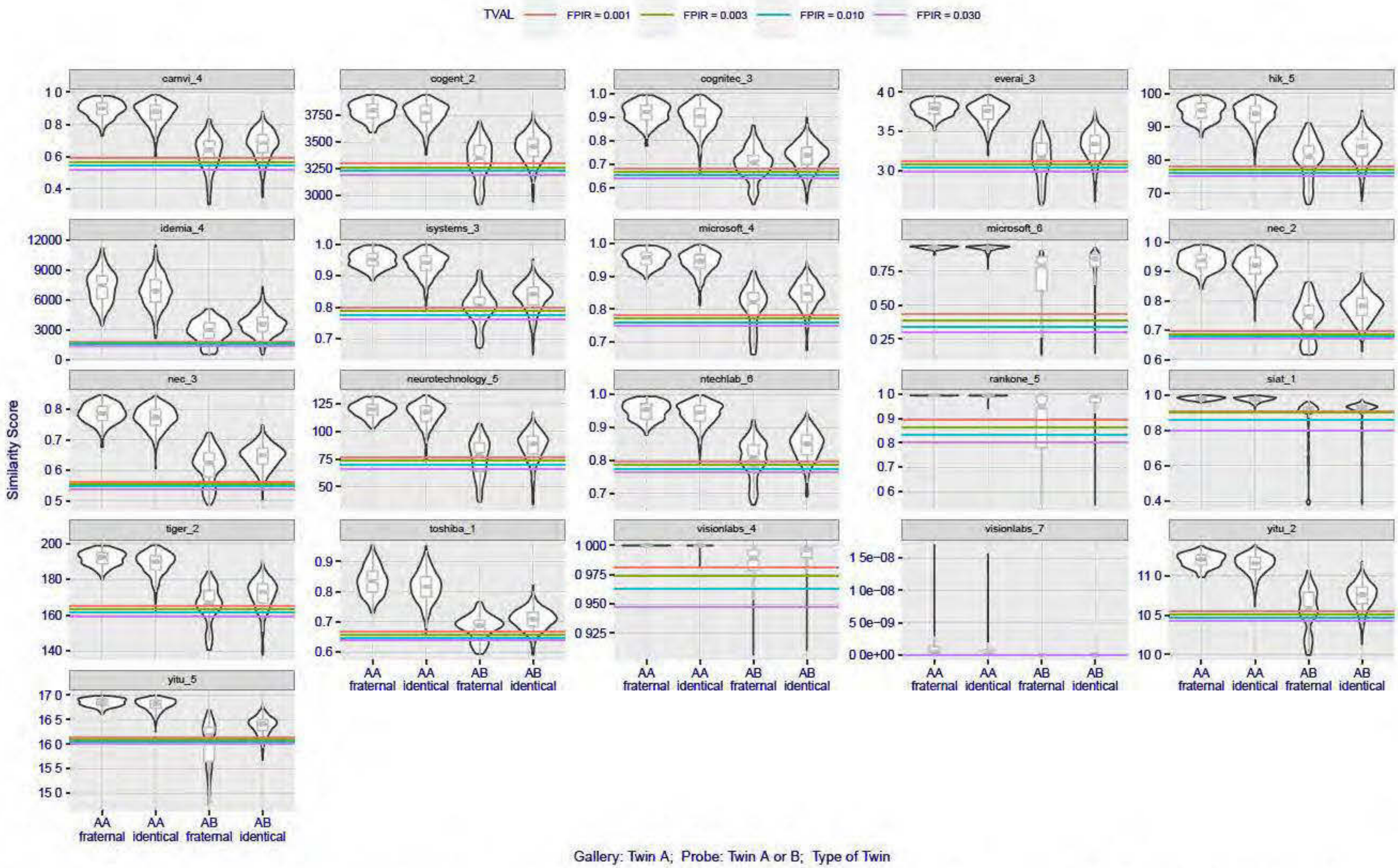


Figure 21: **[Twins Dataset] High scores from twins.** *The Figure shows native similarity scores from searches into a dataset of N = 640 000 background mugshot images plus 104 portrait images, one from each of one of a pair of twins. Two distributions of scores are plotted for each of monozygotic (identical) and dizygotic (fraternal) twins. The first distribution ("AA") shows the mate score from Twin A against their own enrollment. The second ("AB") shows scores from searches of Twin B against the Twin A enrollment: As these are non-mate scores they should be below the various thresholds shown as horizontal lines. That they usually are not is an indication that twins produce very high non-mate scores. Note in theory half of dizygotic (fraternal) twins are different sex. In the sample used here some fraternal twins are correctly rejected.*

Appendices

Appendix A Accuracy on large-population FRVT 2018 mugshots

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

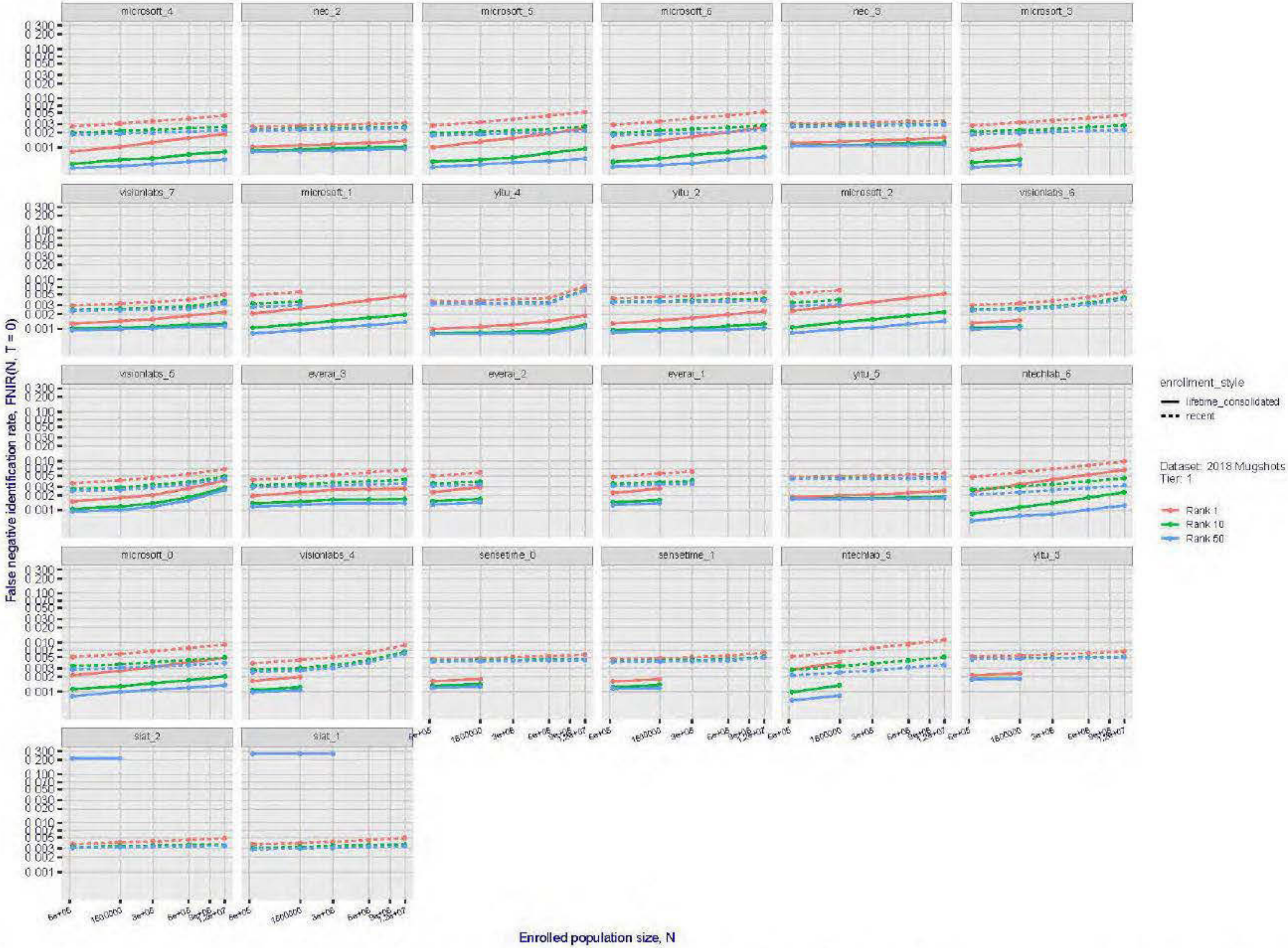


Figure 22: **[FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. number of enrolled subjects.** *The figure shows false negative identification rates, FNIR(N, R), across various gallery sizes and ranks 1, 10 and 50. The threshold is set to zero, so this metric rewards even weak scoring rank 1 mates. This also means FPIR = 1, so any search without an enrolled mate will return non-mated candidates. For clarity, results are sorted and reported into tiers spanning multiple pages, the tiering criteria being rank 1 hit rate on a gallery size of 640000.*

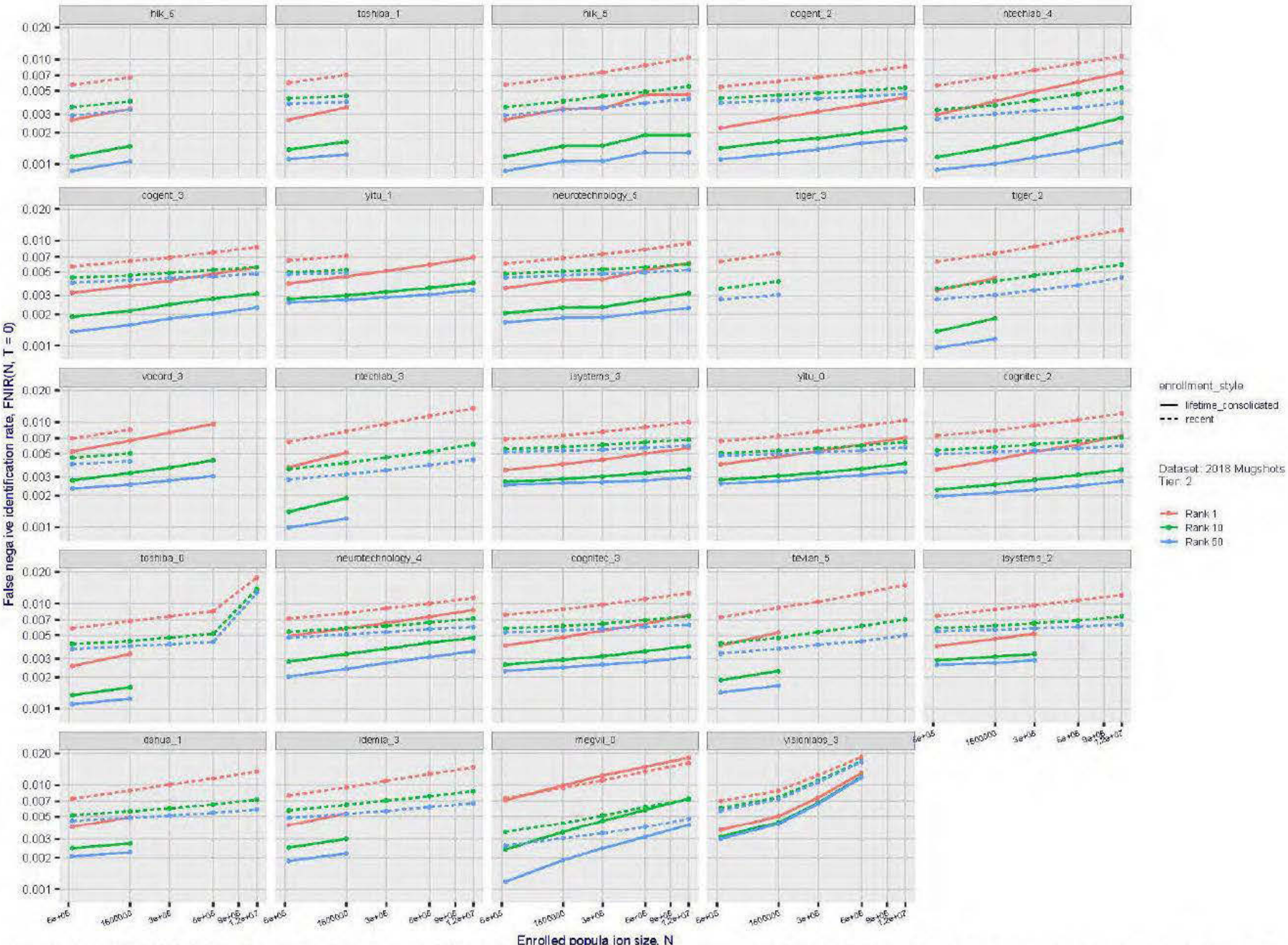


Figure 23: **[FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. number of enrolled subjects.** *The figure shows false negative identification rates, FNIR(N, R), across various gallery sizes and ranks 1, 10 and 50. The threshold is set to zero, so this metric rewards even weak scoring rank 1 mates. This also means FPIR = 1, so any search without an enrolled mate will return non-mated candidates. For clarity, results are sorted and reported into tiers spanning multiple pages, the tiering criteria being rank 1 hit rate on a gallery size of 640 000.*

2019/09/11
17:24:52

FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

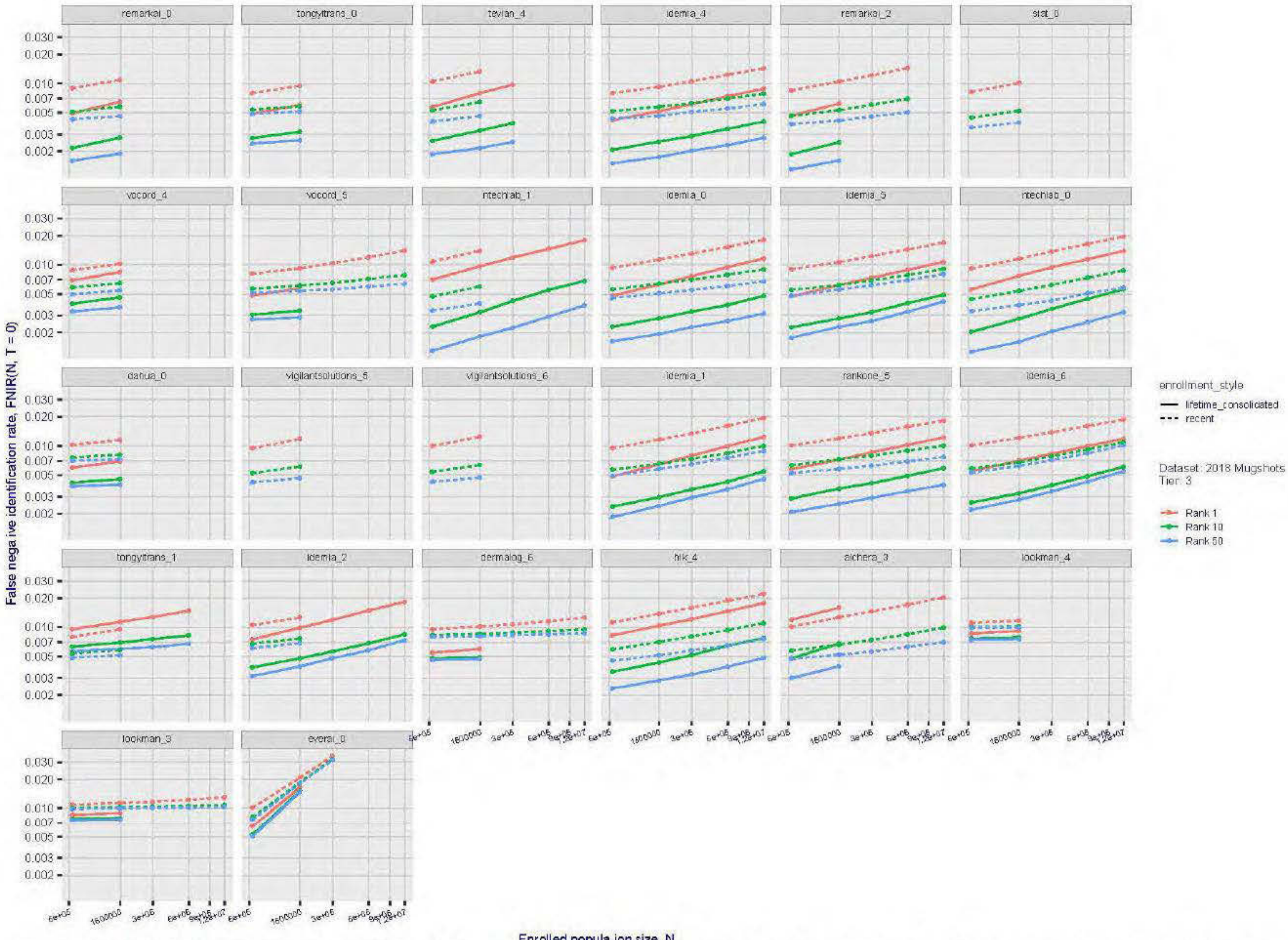


Figure 24: **[FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. number of enrolled subjects.** *The figure shows false negative identification rates, FNIR(N, R), across various gallery sizes and ranks 1, 10 and 50. The threshold is set to zero, so this metric rewards even weak scoring rank 1 mates. This also means FPIR = 1, so any search without an enrolled mate will return non-mated candidates. For clarity, results are sorted and reported into tiers spanning multiple pages, the tiering criteria being rank 1 hit rate on a gallery size of 640 000.*

FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

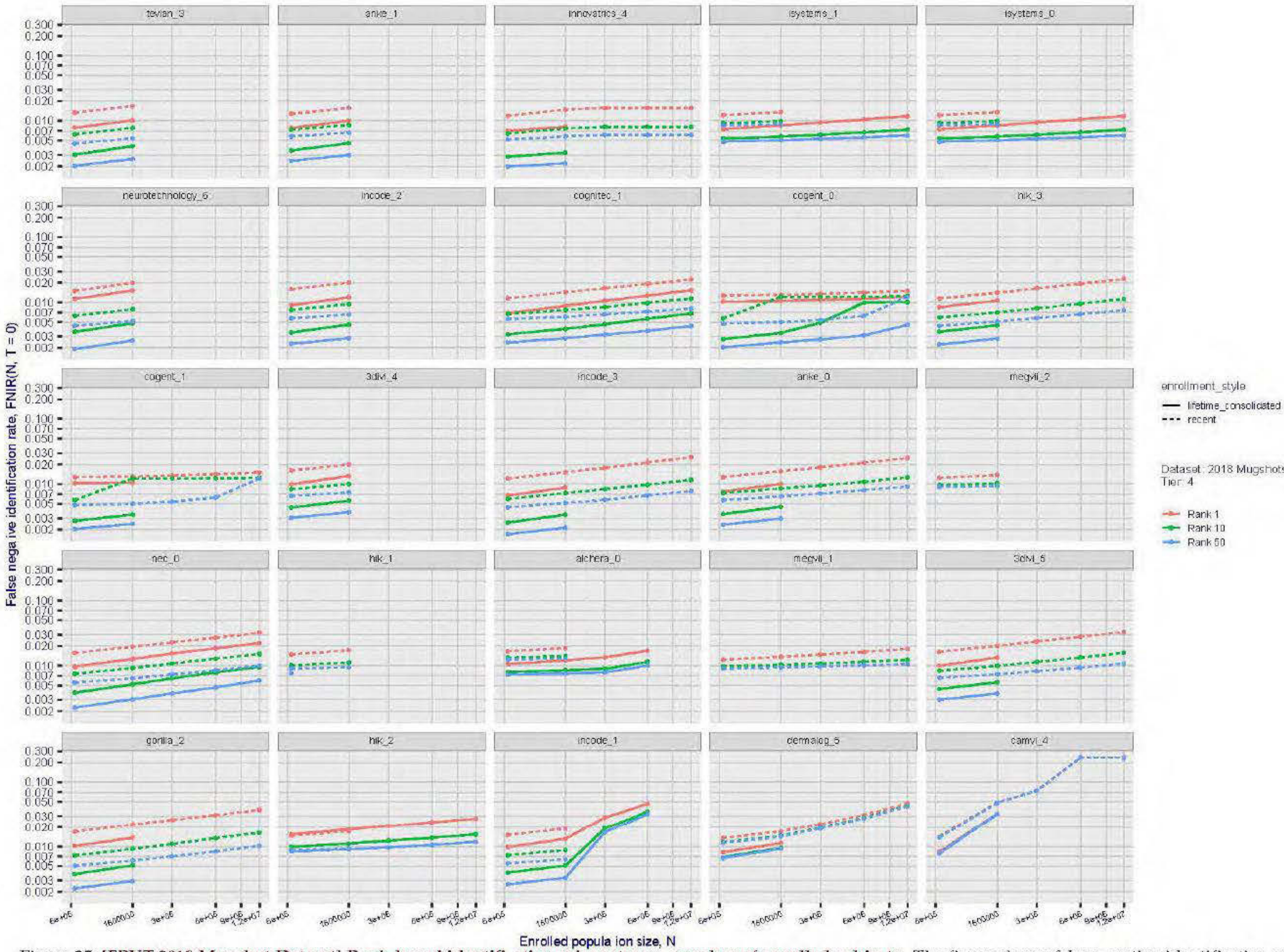


Figure 25: **[FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. number of enrolled subjects.** *The figure shows false negative identification rates, FNIR(N, R), across various gallery sizes and ranks 1, 10 and 50. The threshold is set to zero, so this metric rewards even weak scoring rank 1 mates. This also means FPIR = 1, so any search without an enrolled mate will return non-mated candidates. For clarity, results are sorted and reported into tiers spanning multiple pages, the tiering criteria being rank 1 hit rate on a gallery size of 640 000.*

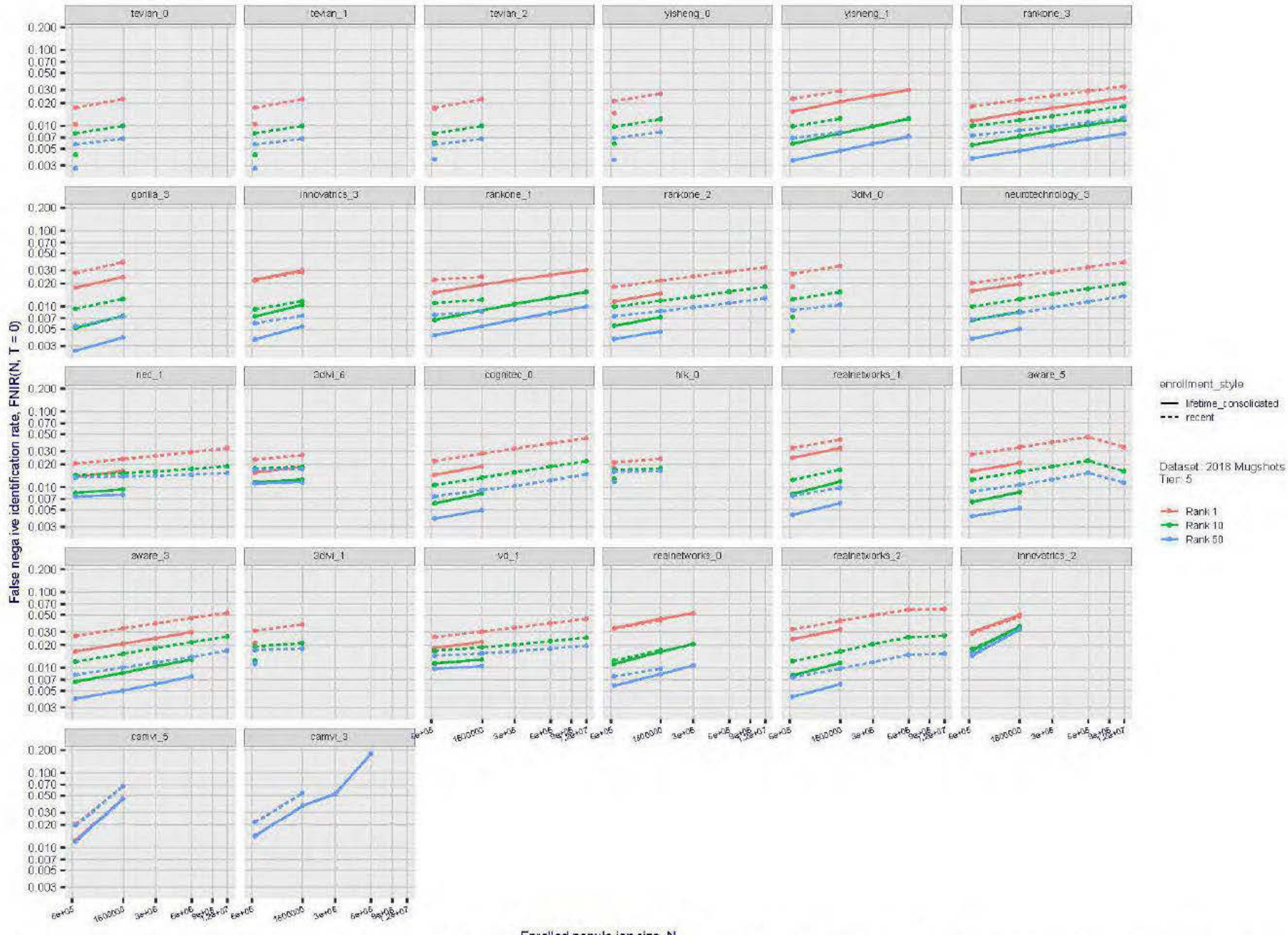


Figure 26: **[FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. number of enrolled subjects.** *The figure shows false negative identification rates, FNIR(N, R), across various gallery sizes and ranks 1, 10 and 50. The threshold is set to zero, so this metric rewards even weak scoring rank 1 mates. This also means FPIR = 1, so any search without an enrolled mate will return non-mated candidates. For clarity, results are sorted and reported into tiers spanning multiple pages, the tiering criteria being rank 1 hit rate on a gallery size of 640 000.*

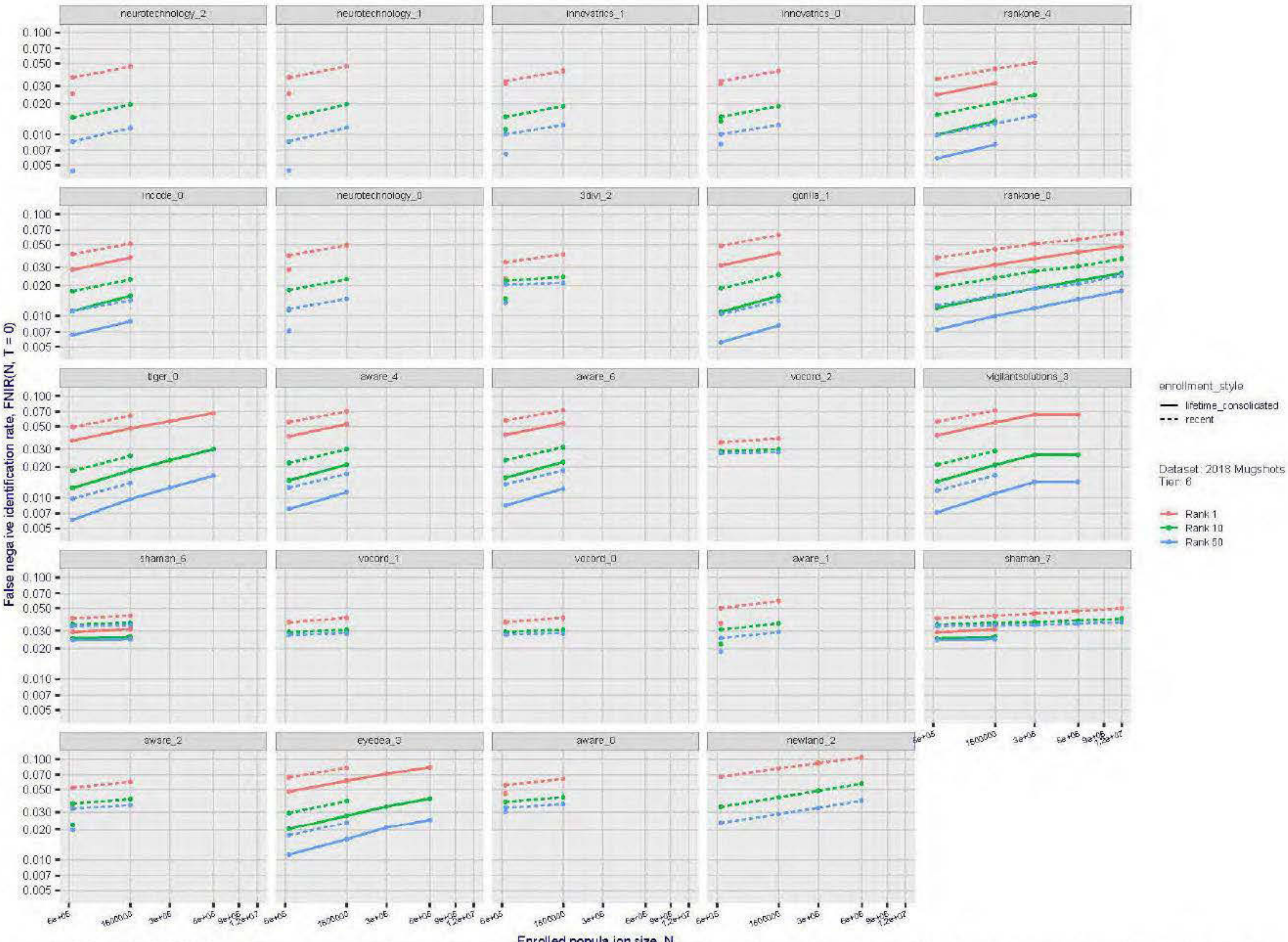


Figure 27: **[FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. number of enrolled subjects.** *The figure shows false negative identification rates, FNIR(N, R), across various gallery sizes and ranks 1, 10 and 50. The threshold is set to zero, so this metric rewards even weak scoring rank 1 mates. This also means FPIR = 1, so any search without an enrolled mate will return non-mated candidates. For clarity, results are sorted and reported into tiers spanning multiple pages, the tiering criteria being rank 1 hit rate on a gallery size of 640 000.*

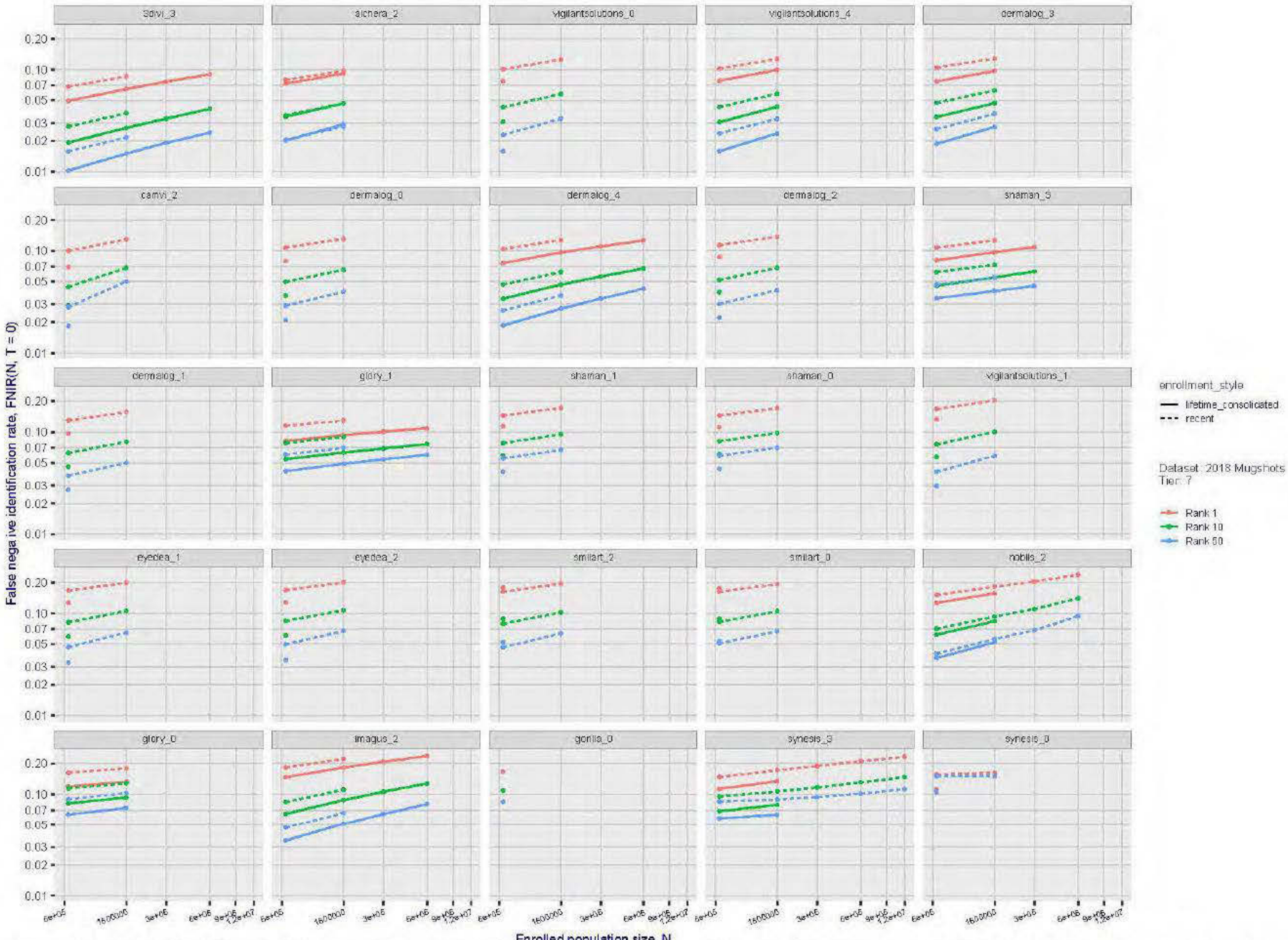


Figure 28: **[FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. number of enrolled subjects.** *The figure shows false negative identification rates, FNIR(N, R), across various gallery sizes and ranks 1, 10 and 50. The threshold is set to zero, so this metric rewards even weak scoring rank 1 mates. This also means FPIR = 1, so any search without an enrolled mate will return non-mated candidates. For clarity, results are sorted and reported into tiers spanning multiple pages, the tiering criteria being rank 1 hit rate on a gallery size of 640 000.*

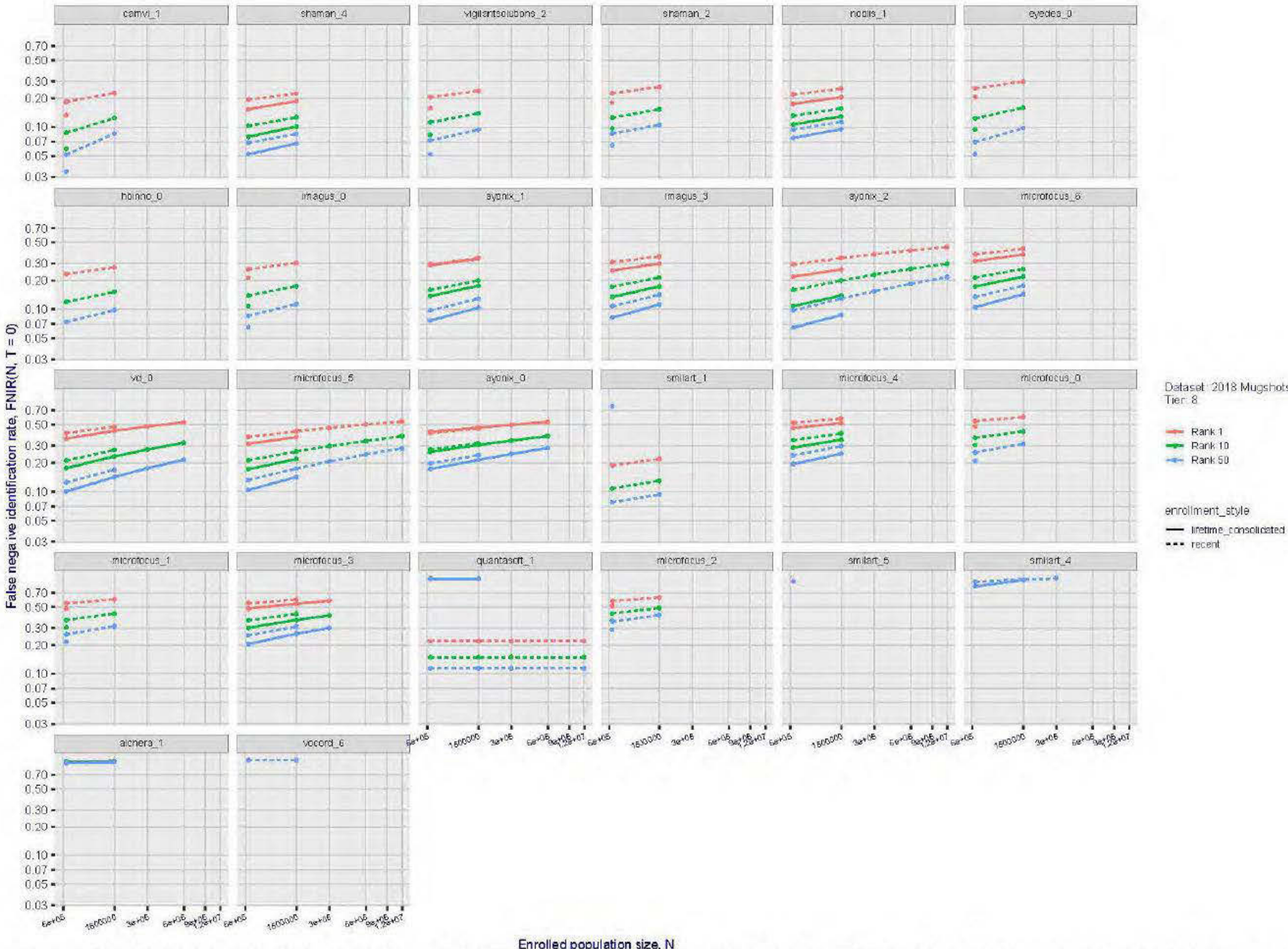


Figure 29: **[FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. number of enrolled subjects.** *The figure shows false negative identification rates, FNIR(N, R), across various gallery sizes and ranks 1, 10 and 50. The threshold is set to zero, so this metric rewards even weak scoring rank 1 mates. This also means FPIR = 1, so any search without an enrolled mate will return non-mated candidates. For clarity, results are sorted and reported into tiers spanning multiple pages, the tiering criteria being rank 1 hit rate on a gallery size of 640 000.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

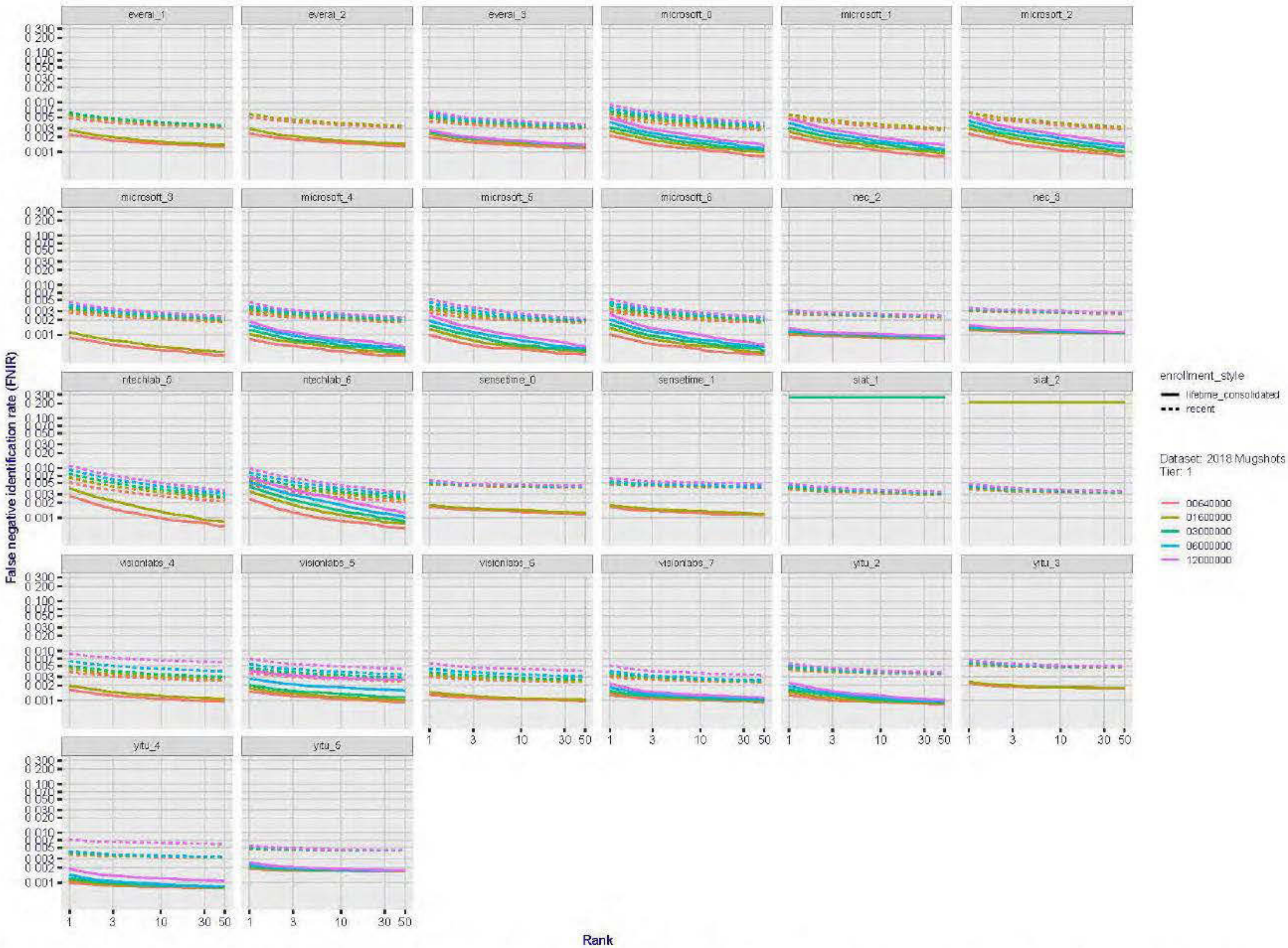


Figure 30: **[FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. rank.** *The figure shows false negative identification rates (FNIR) for ranks up to 50. This metric is appropriate to investigational applications where human reviewers will adjudicate sorted candidate lists. Note that with threshold set to zero, FPIR = 1, i.e. any search without an enrolled mate will return non-mated candidates. Results are sorted and reported into tiers for clarity, with the tiering criteria being rank 1 hit rate on a gallery size of N = 640 000 subjects.*

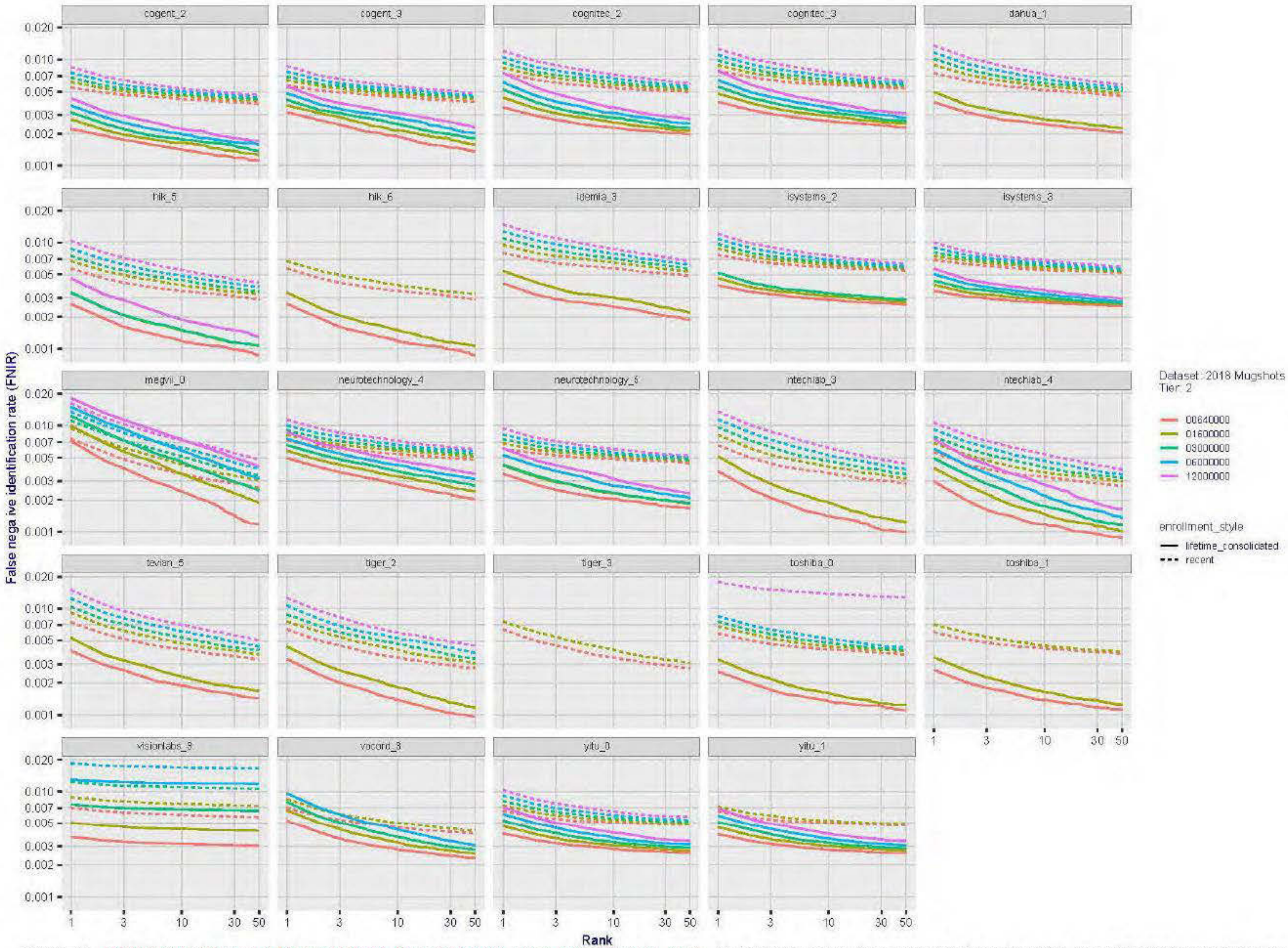


Figure 31: **[FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. rank.** *The figure shows false negative identification rates (FNIR) for ranks up to 50. This metric is appropriate to investigational applications where human reviewers will adjudicate sorted candidate lists. Note that with threshold set to zero, FPIR = 1, i.e. any search without an enrolled mate will return non-mated candidates. Results are sorted and reported into tiers for clarity, with the tiering criteria being rank 1 hit rate on a gallery size of N = 640 000 subjects.*

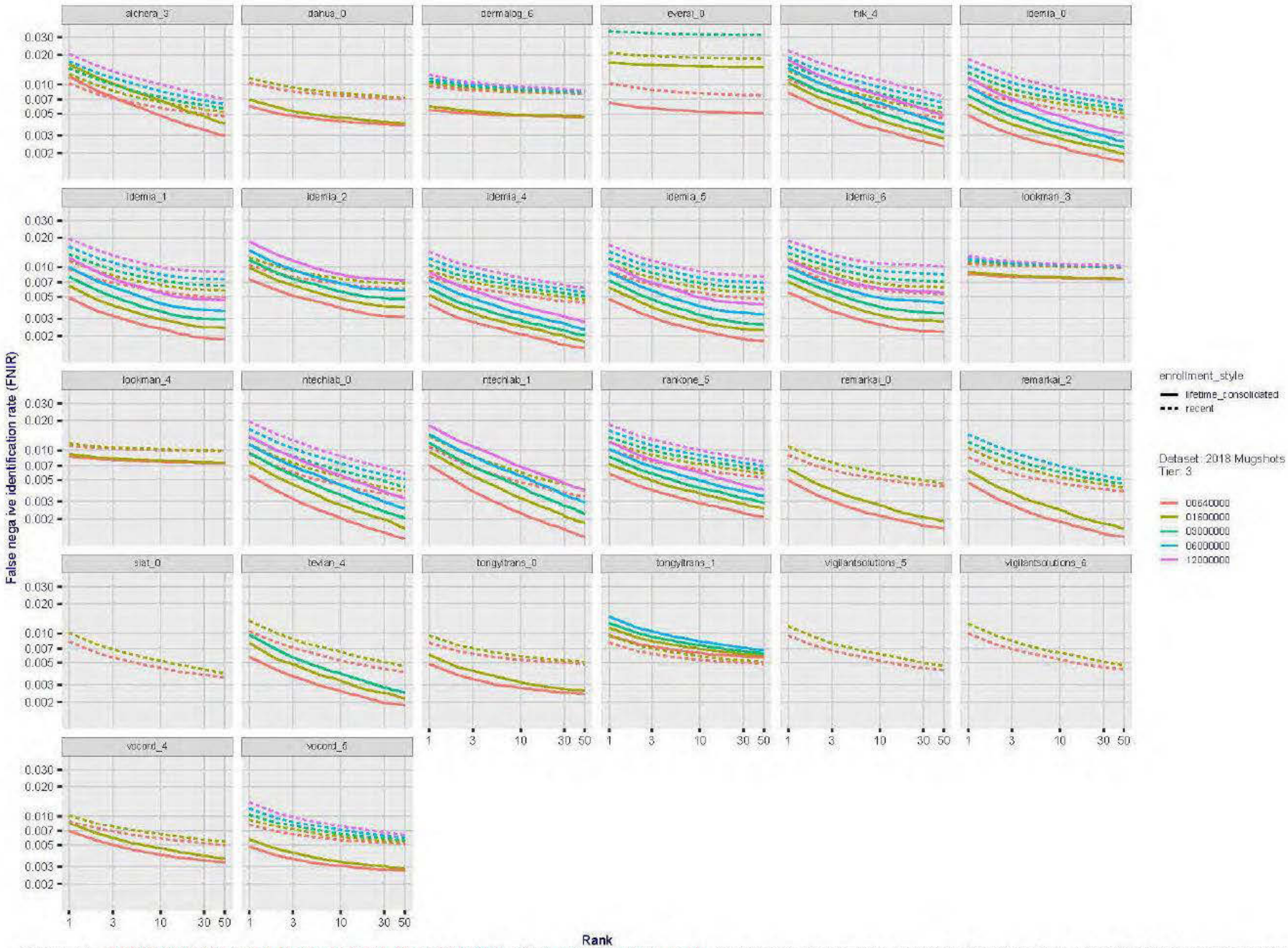


Figure 32: **[FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. rank.** *The figure shows false negative identification rates (FNIR) for ranks up to 50. This metric is appropriate to investigational applications where human reviewers will adjudicate sorted candidate lists. Note that with threshold set to zero, FPIR = 1, i.e. any search without an enrolled mate will return non-mated candidates. Results are sorted and reported into tiers for clarity, with the tiering criteria being rank 1 hit rate on a gallery size of N = 640 000 subjects.*

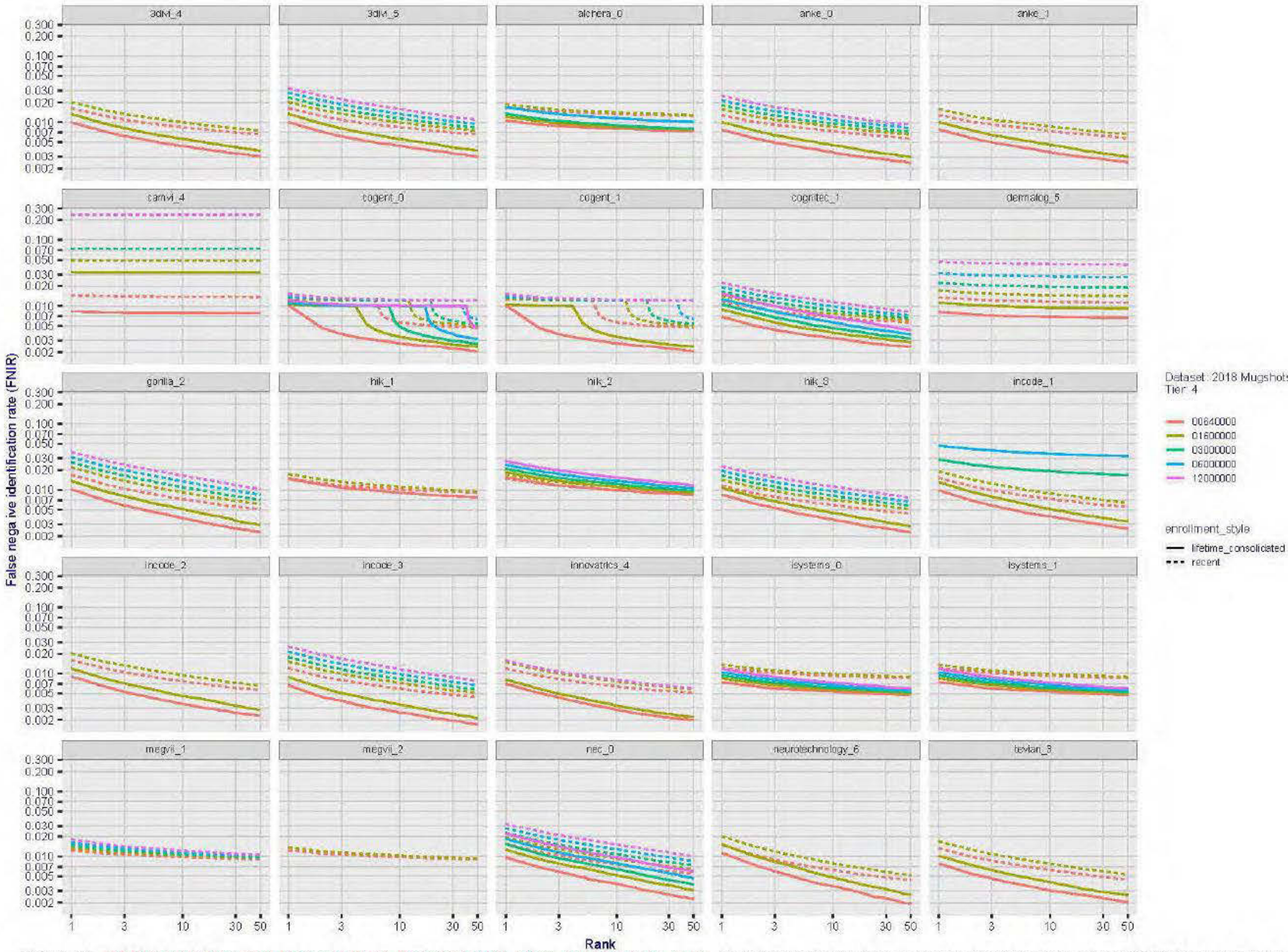


Figure 33: **[FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. rank.** *The figure shows false negative identification rates (FNIR) for ranks up to 50. This metric is appropriate to investigational applications where human reviewers will adjudicate sorted candidate lists. Note that with threshold set to zero, FPIR = 1, i.e. any search without an enrolled mate will return non-mated candidates. Results are sorted and reported into tiers for clarity, with the tiering criteria being rank 1 hit rate on a gallery size of N = 640 000 subjects.*

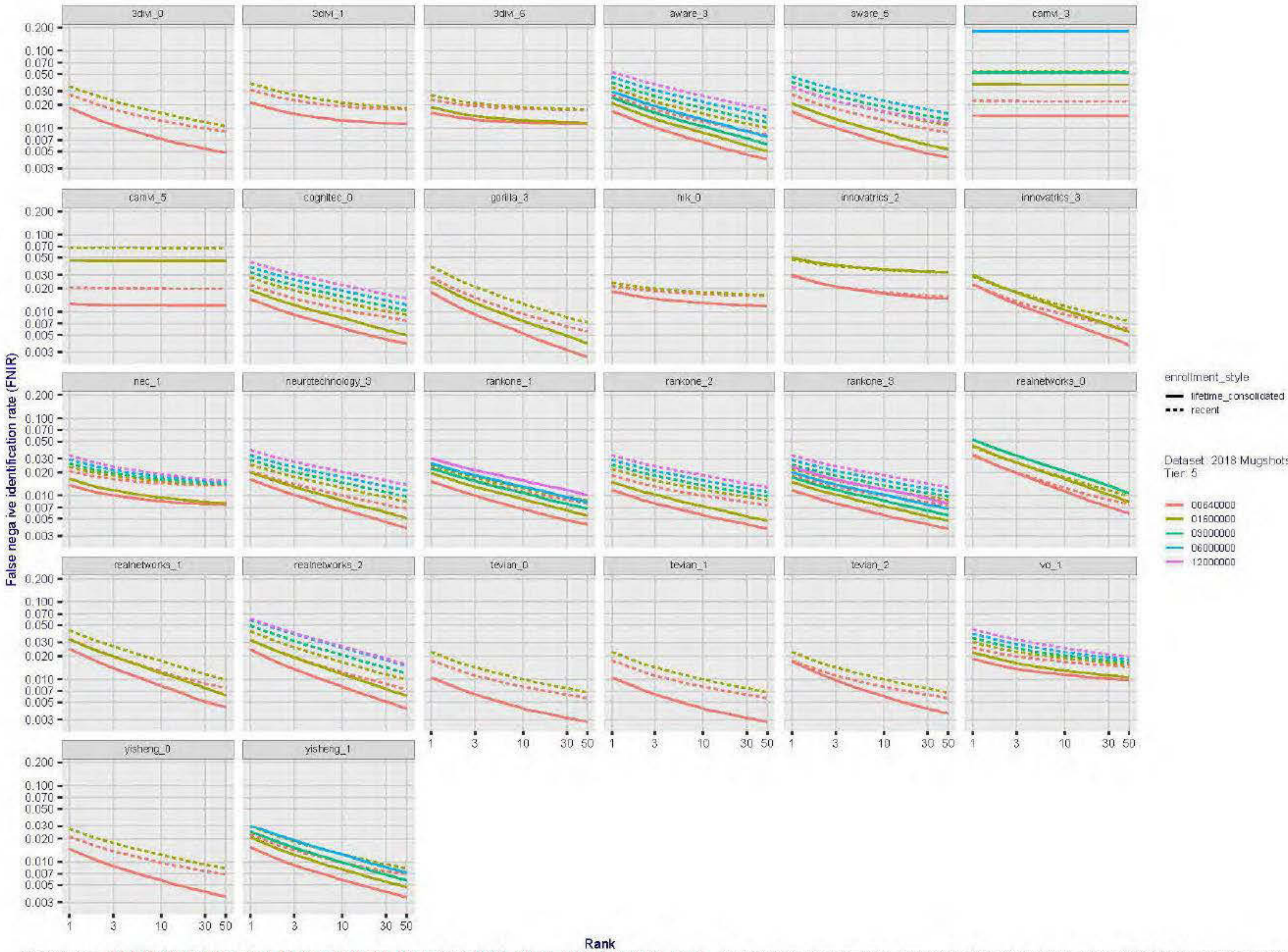


Figure 34: **[FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. rank.** *The figure shows false negative identification rates (FNIR) for ranks up to 50. This metric is appropriate to investigational applications where human reviewers will adjudicate sorted candidate lists. Note that with threshold set to zero, FPIR = 1, i.e. any search without an enrolled mate will return non-mated candidates. Results are sorted and reported into tiers for clarity, with the tiering criteria being rank 1 hit rate on a gallery size of N = 640 000 subjects.*

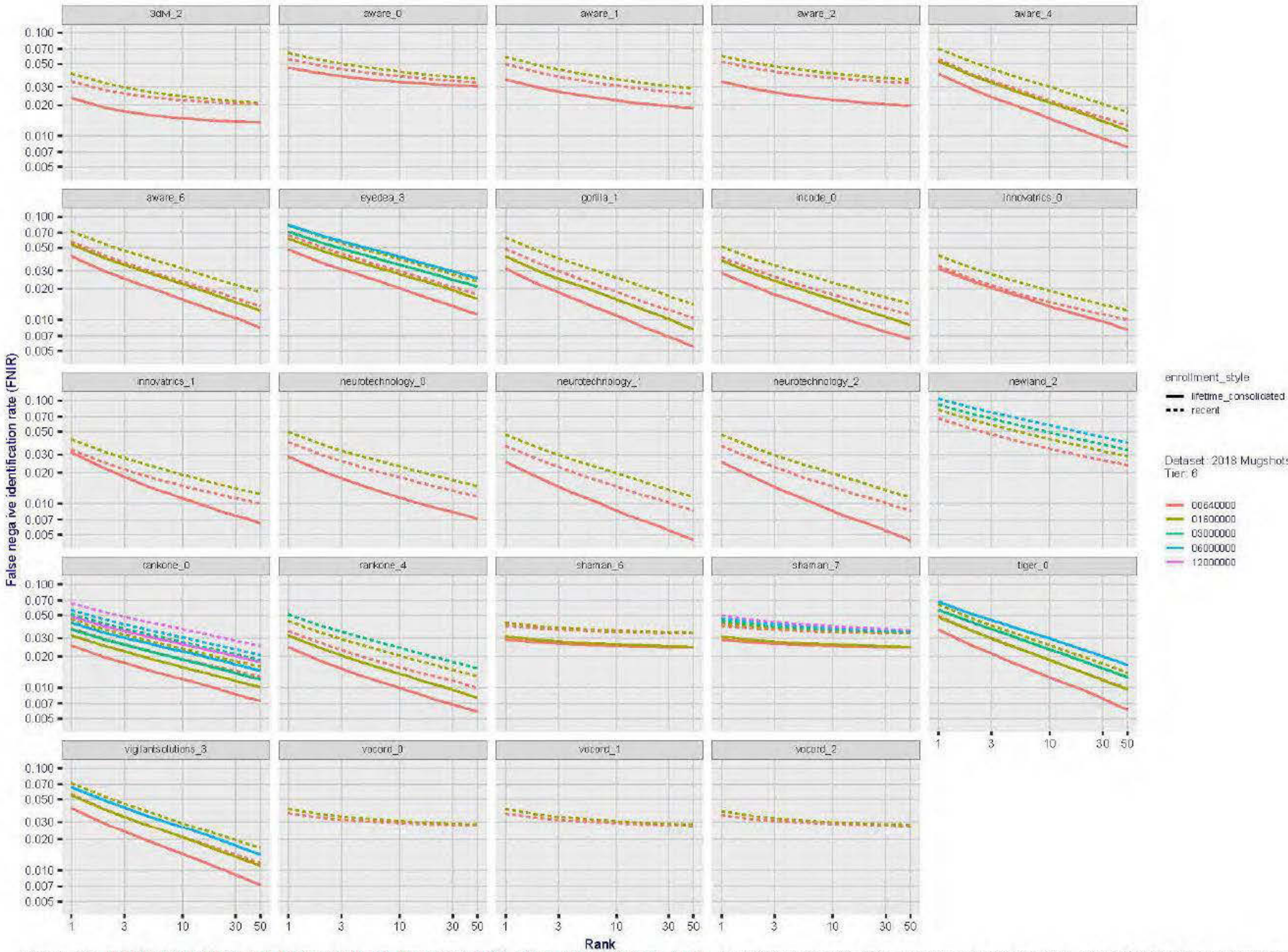


Figure 35: **[FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. rank.** *The figure shows false negative identification rates (FNIR) for ranks up to 50. This metric is appropriate to investigational applications where human reviewers will adjudicate sorted candidate lists. Note that with threshold set to zero, FPIR = 1, i.e. any search without an enrolled mate will return non-mated candidates. Results are sorted and reported into tiers for clarity, with the tiering criteria being rank 1 hit rate on a gallery size of N = 640 000 subjects.*

FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

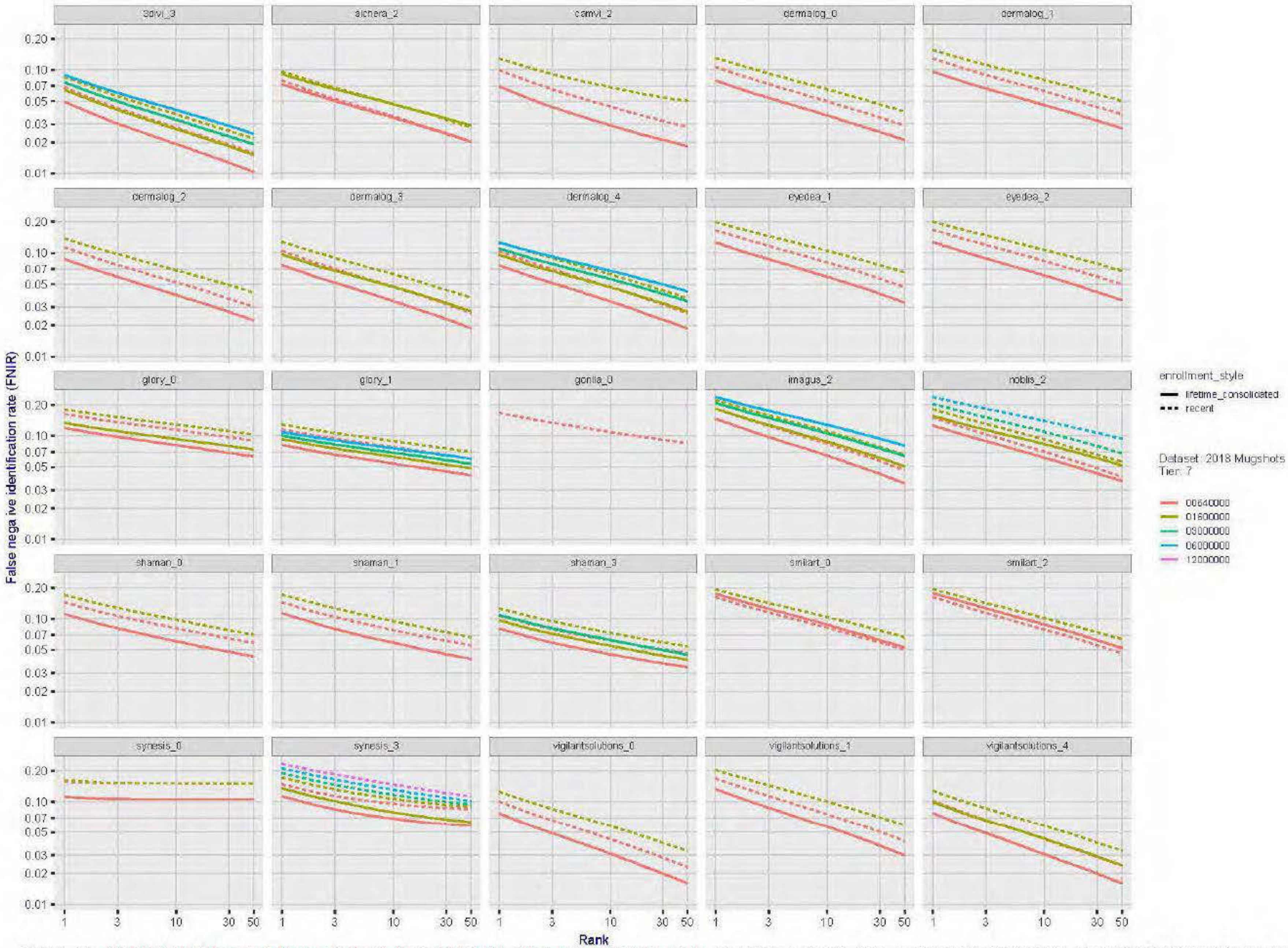


Figure 36: **[FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. rank.** *The figure shows false negative identification rates (FNIR) for ranks up to 50. This metric is appropriate to investigational applications where human reviewers will adjudicate sorted candidate lists. Note that with threshold set to zero, FPIR = 1, i.e. any search without an enrolled mate will return non-mated candidates. Results are sorted and reported into tiers for clarity, with the tiering criteria being rank 1 hit rate on a gallery size of N = 640 000 subjects.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

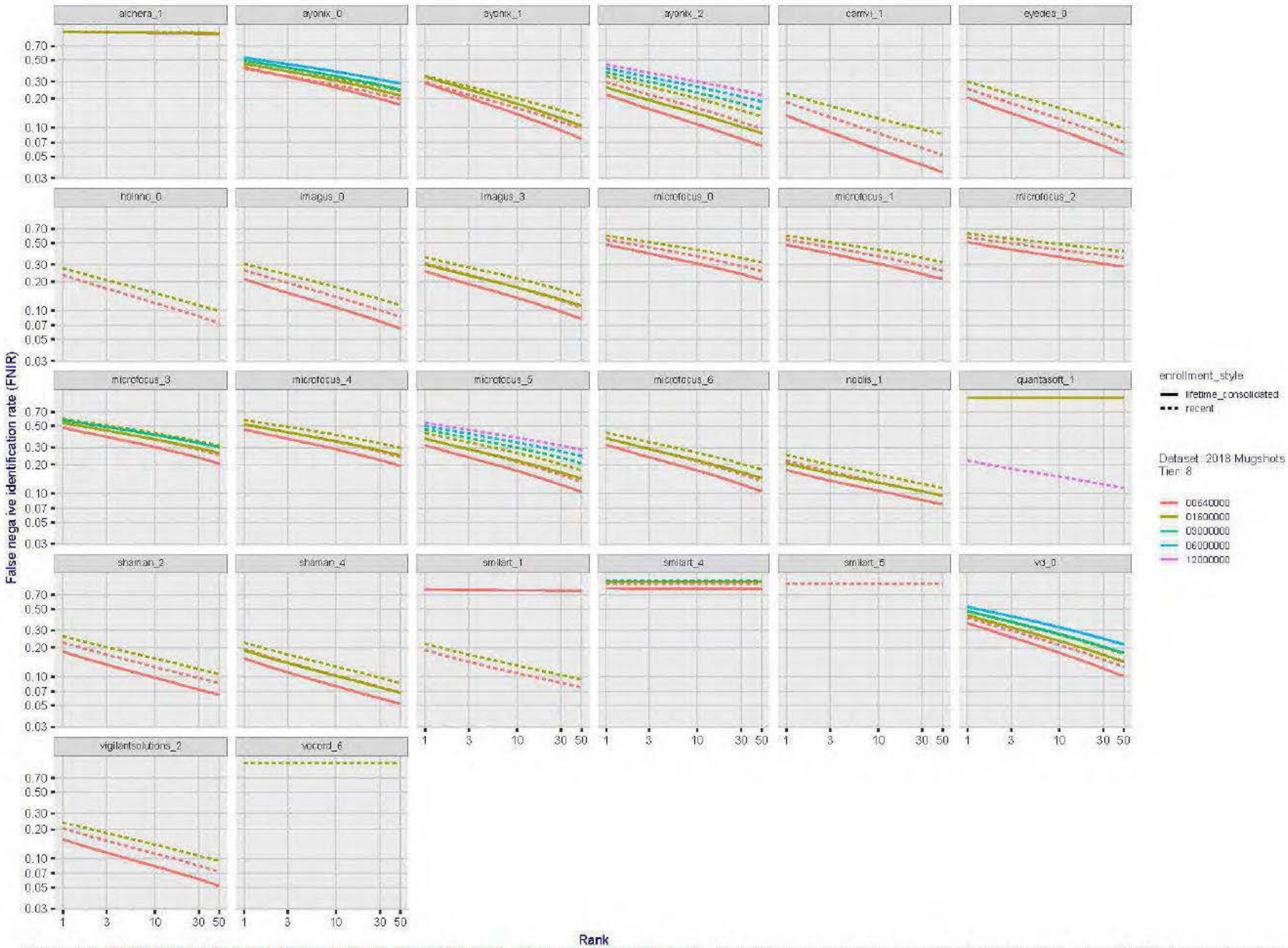


Figure 37: **[FRVT-2018 Mugshot Dataset] Rank-based identification miss rates vs. rank.** *The figure shows false negative identification rates (FNIR) for ranks up to 50. This metric is appropriate to investigational applications where human reviewers will adjudicate sorted candidate lists. Note that with threshold set to zero, FPIR = 1, i.e. any search without an enrolled mate will return non-mated candidates. Results are sorted and reported into tiers for clarity, with the tiering criteria being rank 1 hit rate on a gallery size of N = 640 000 subjects.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

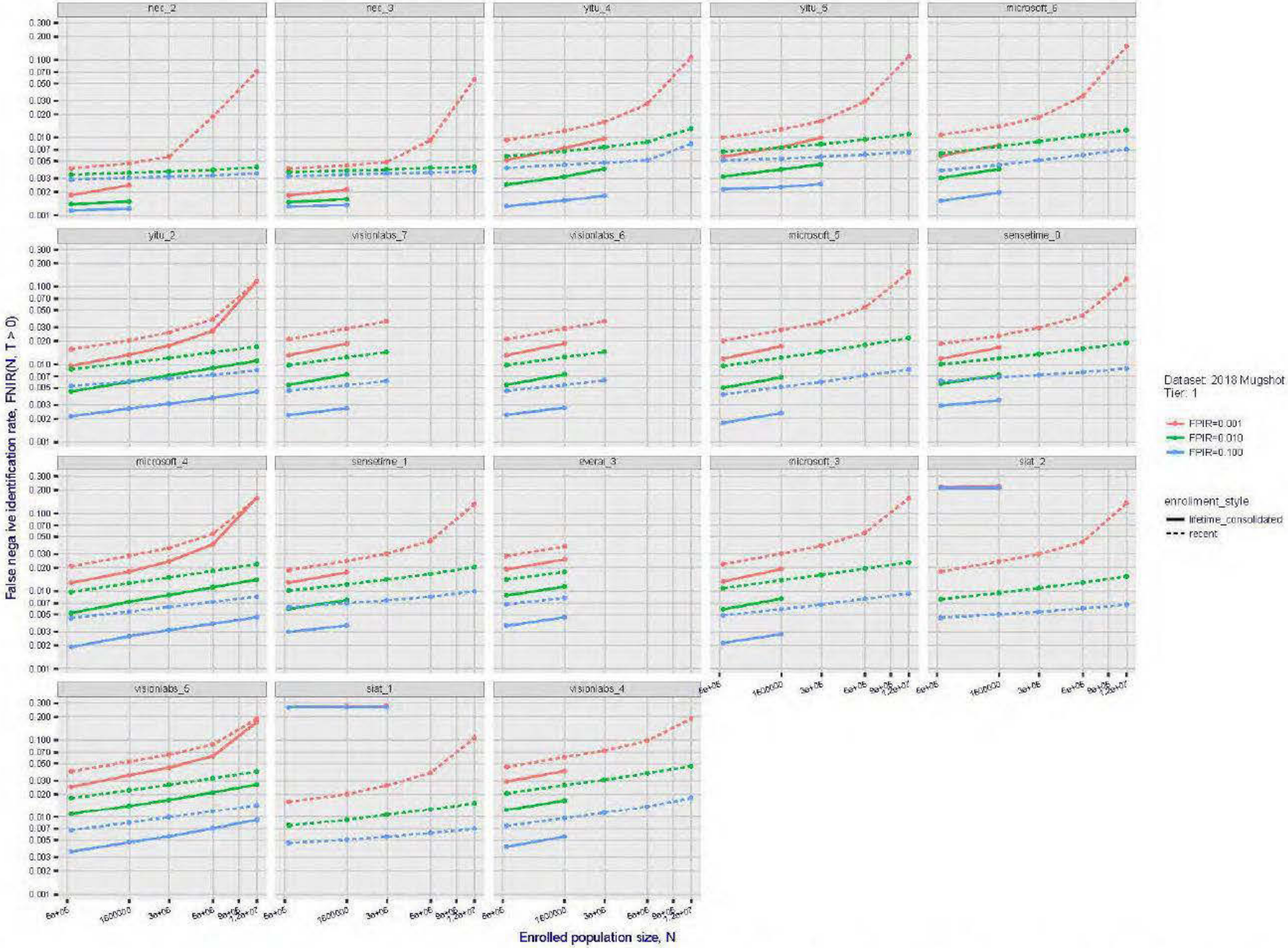


Figure 38: **[FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects.** *The figure shows FNIR(N, T) across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 5. Less accurate algorithms were not run on large N, so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by FNIR(N_b, 1, 0), then sorting by median FNIR(N_b, T), N_b = 640 000.*

FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

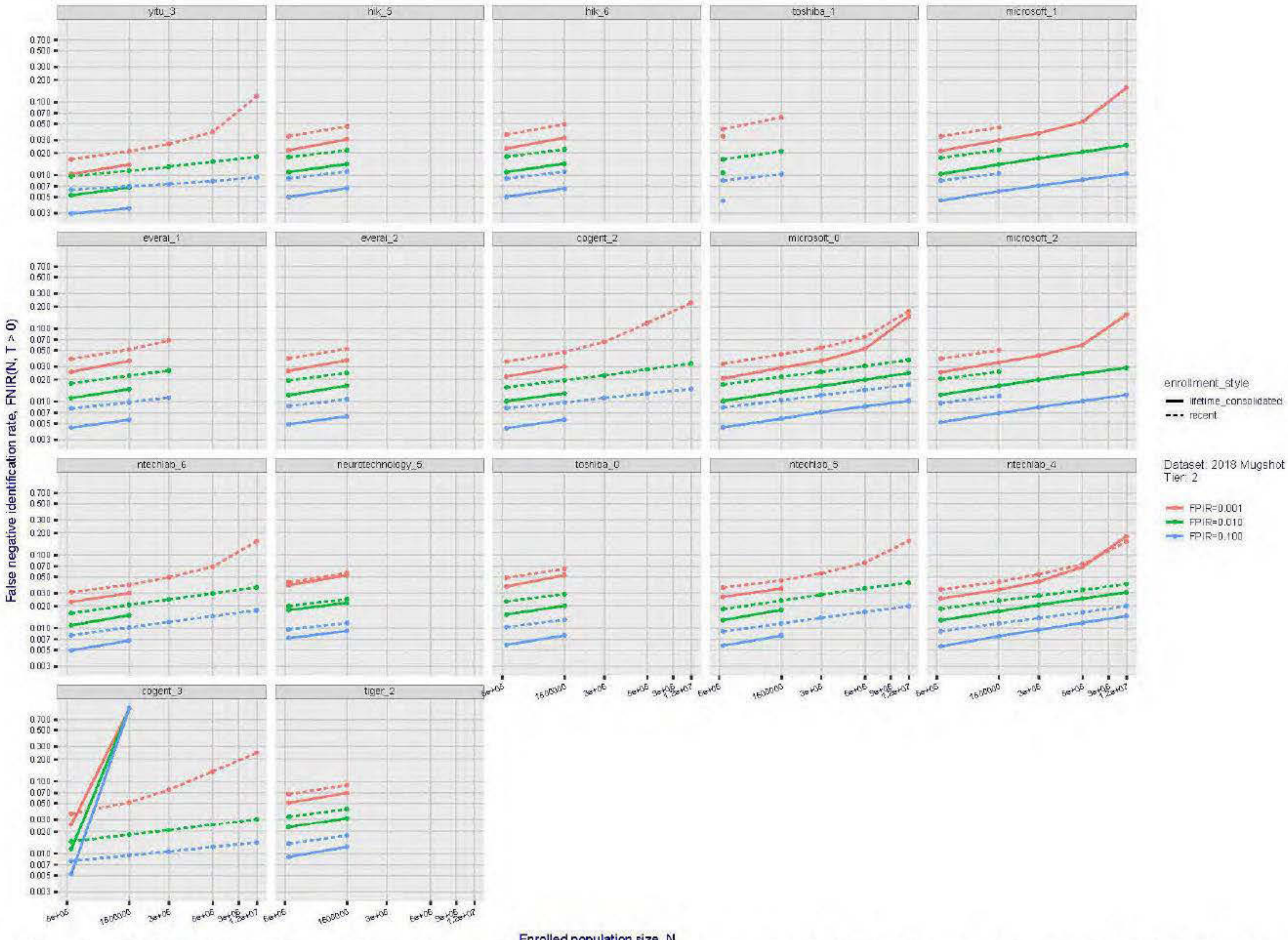


Figure 39: **[FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects.** *The figure shows FNIR(N, T) across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 5. Less accurate algorithms were not run on large N, so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by FNIR(N_b, 1, 0), then sorting by median FNIR(N_b, T), N_b = 640 000.*

FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

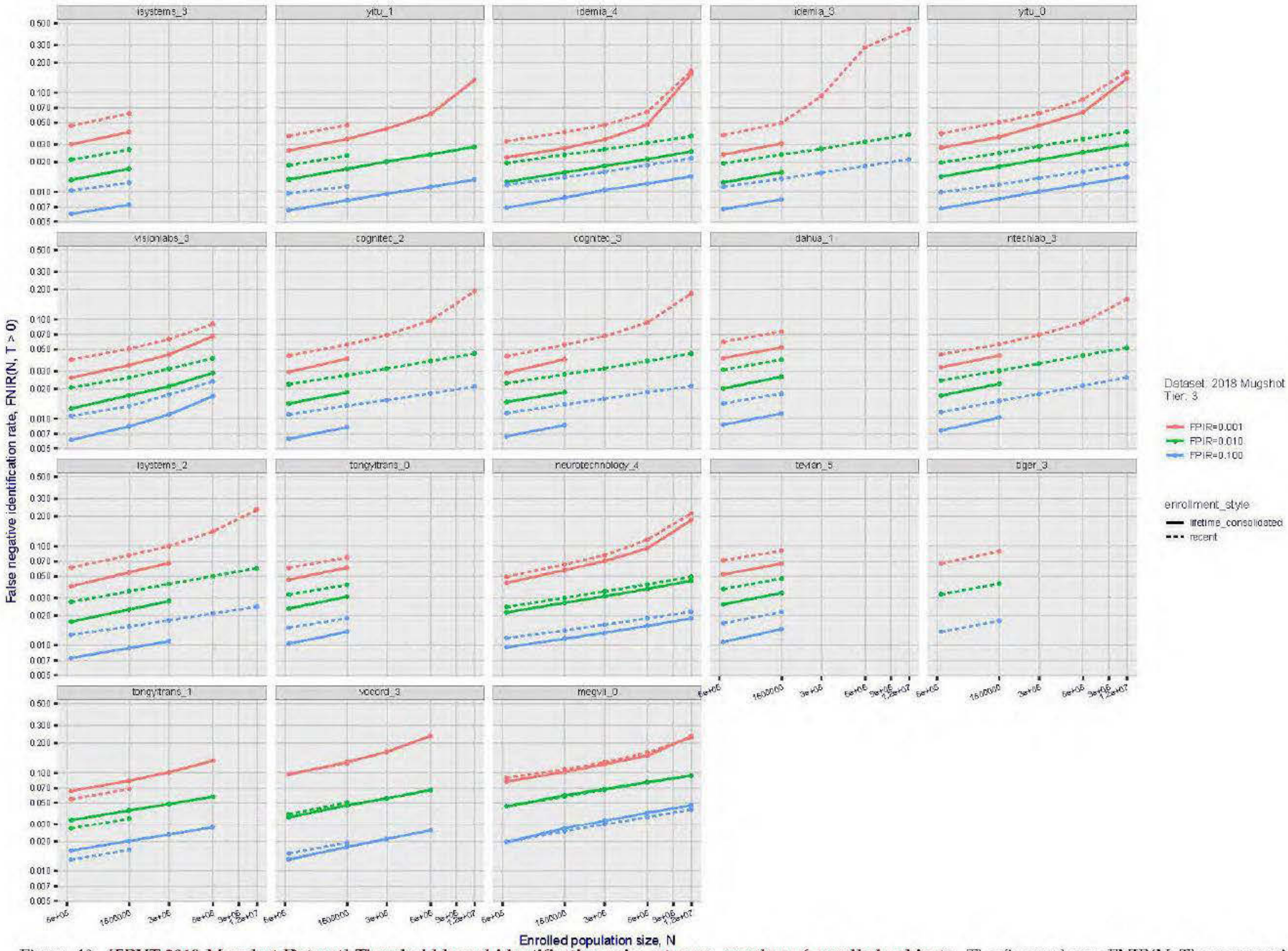


Figure 40: **[FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects.** *The figure shows FNIR(N, T) across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 5. Less accurate algorithms were not run on large N, so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by FNIR(N_b, 1, 0), then sorting by median FNIR(N_b, T), N_b = 640 000.*

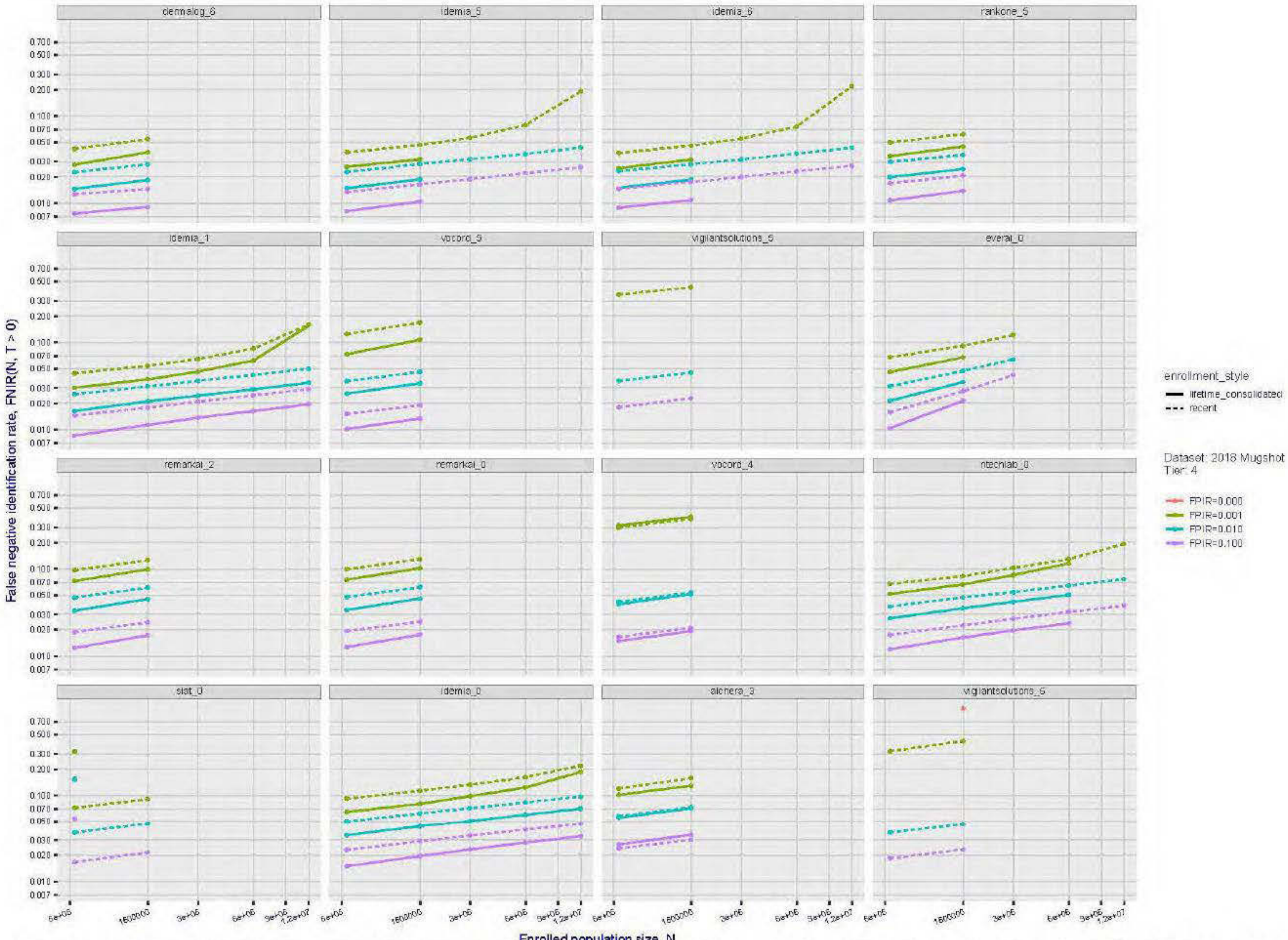


Figure 41: **[FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects.** *The figure shows FNIR(N, T) across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 5. Less accurate algorithms were not run on large N, so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by FNIR(N_b, 1, 0), then sorting by median FNIR(N_b, T), N_b = 640 000.*

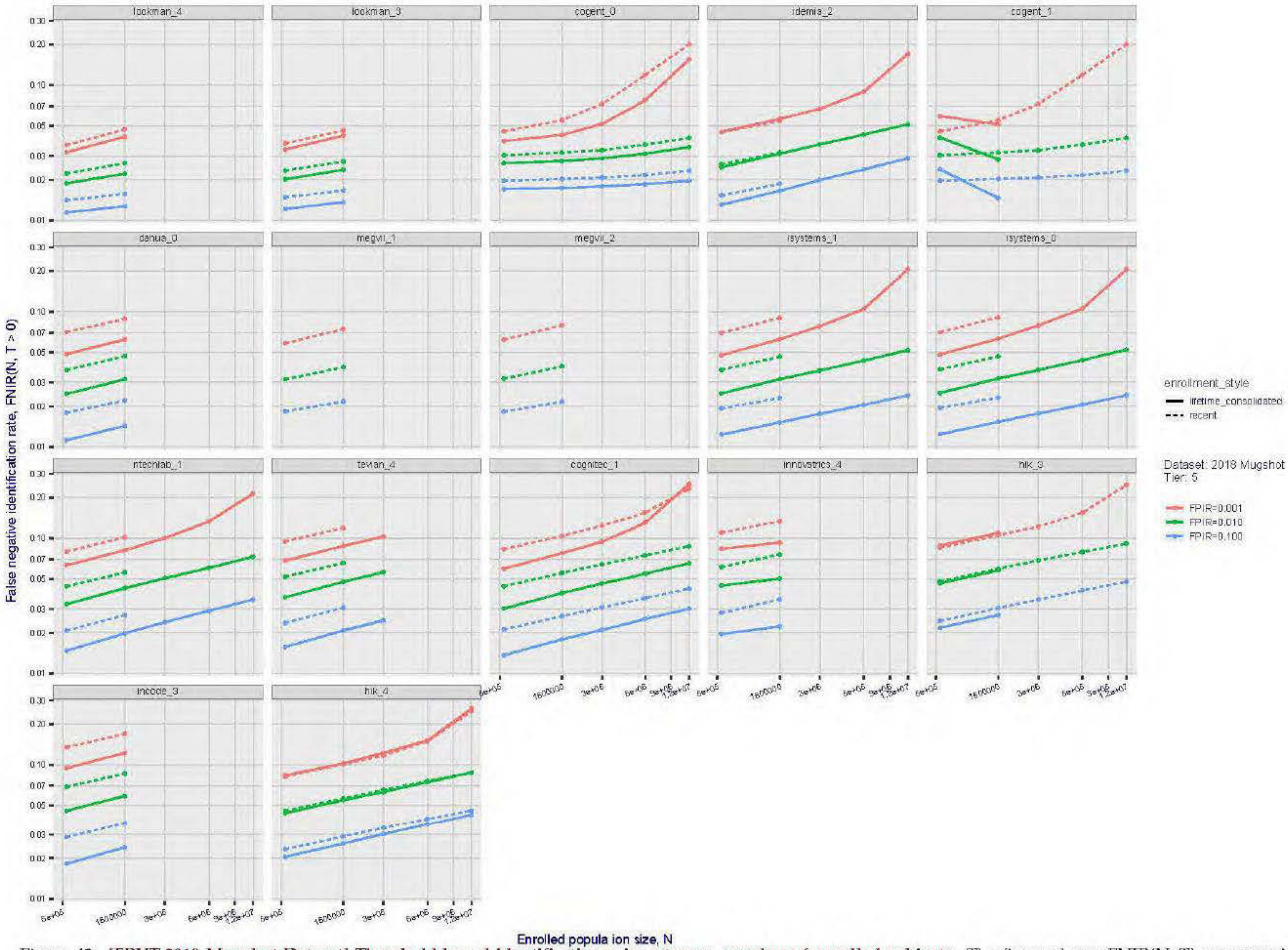


Figure 42: **[FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects.** *The figure shows FNIR(N, T) across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 5. Less accurate algorithms were not run on large N, so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by FNIR(N_b, 1, 0), then sorting by median FNIR(N_b, T), N_b = 640 000.*

FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

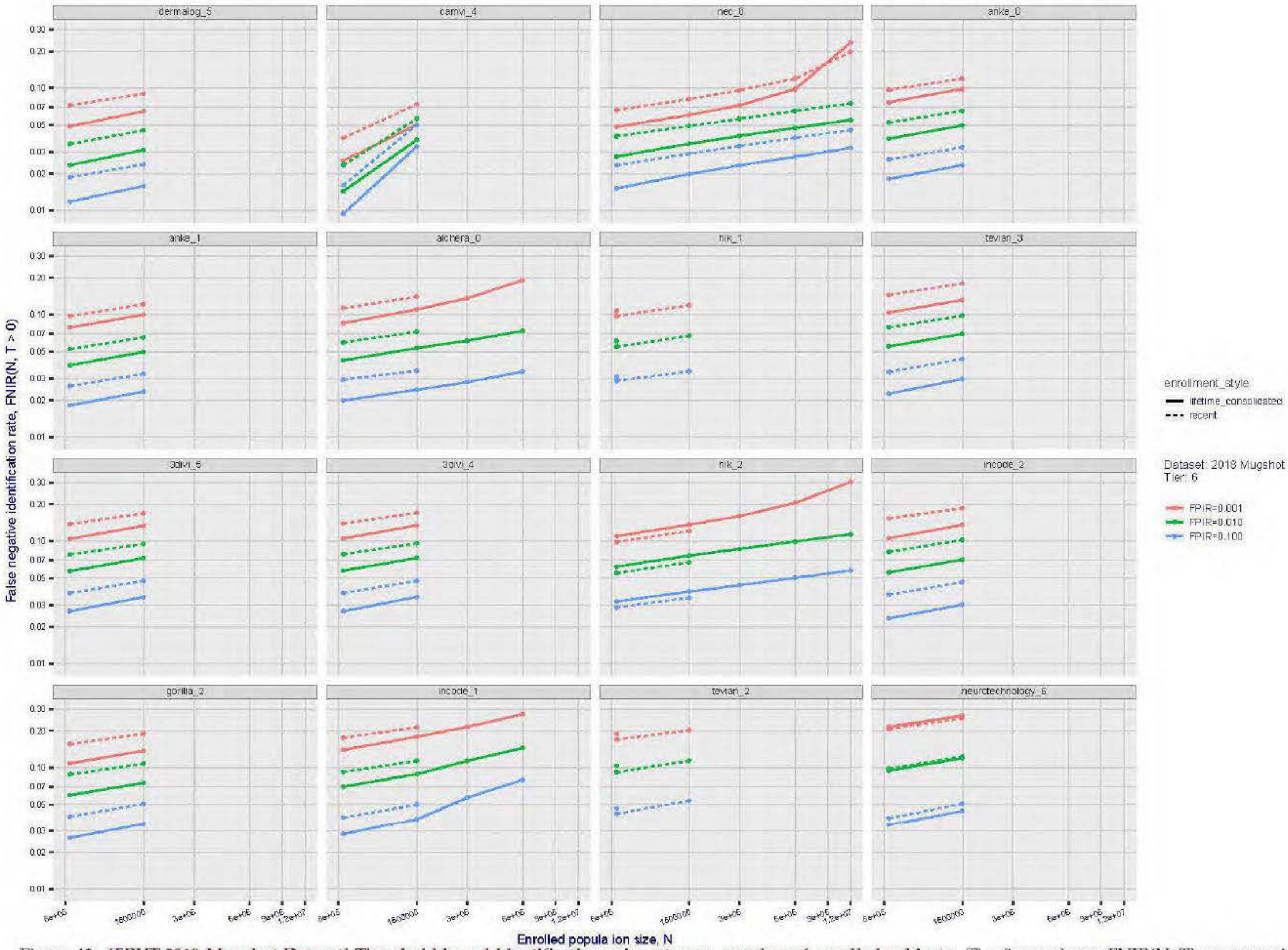


Figure 43: **[FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects.** *The figure shows FNIR(N, T) across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 5. Less accurate algorithms were not run on large N, so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by FNIR(N_b, 1, 0), then sorting by median FNIR(N_b, T), N_b = 640 000.*

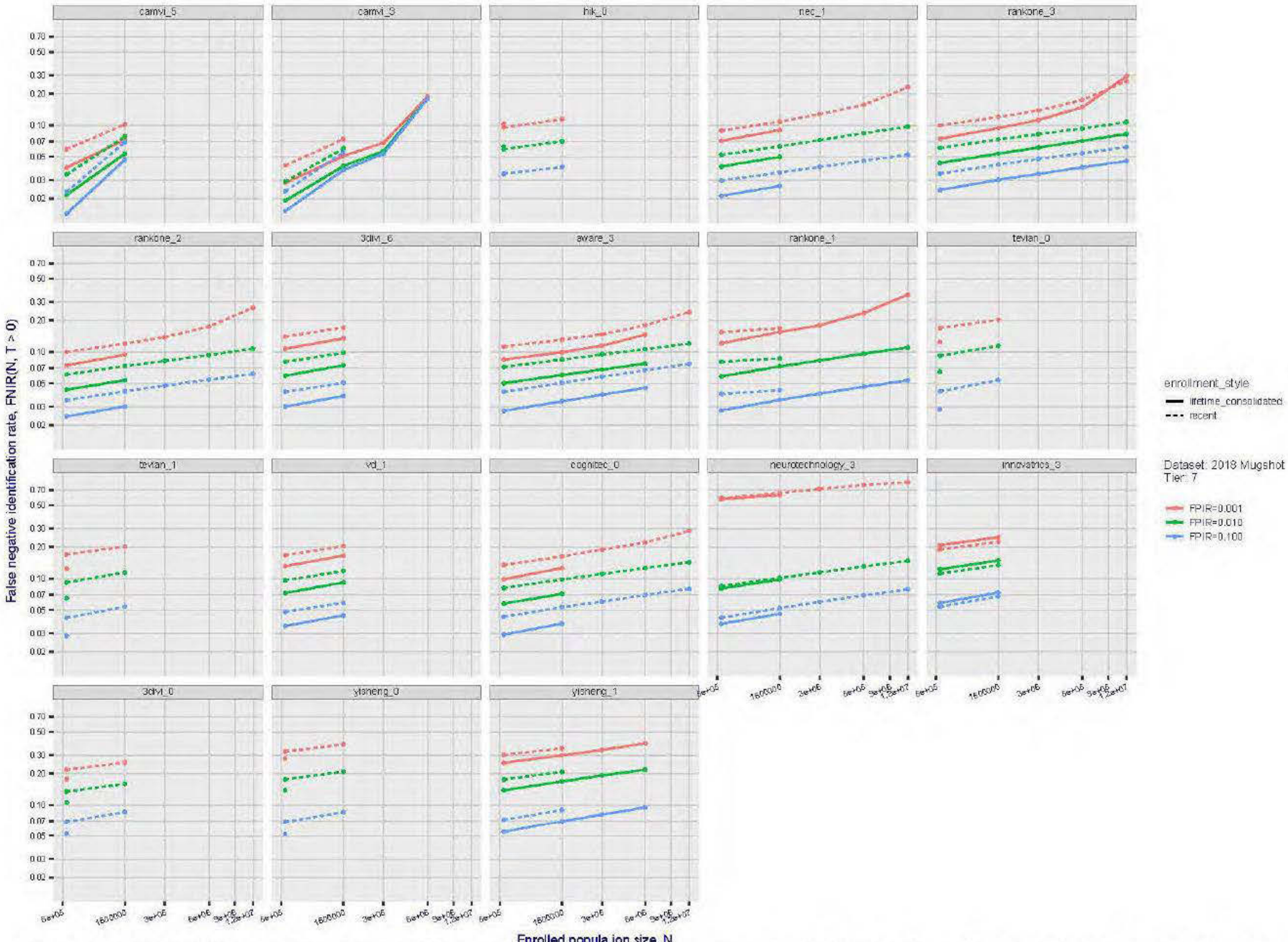


Figure 44: **[FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects.** *The figure shows FNIR(N, T) across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 5. Less accurate algorithms were not run on large N, so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by FNIR(N_b, 1, 0), then sorting by median FNIR(N_b, T), N_b = 640 000.*

FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

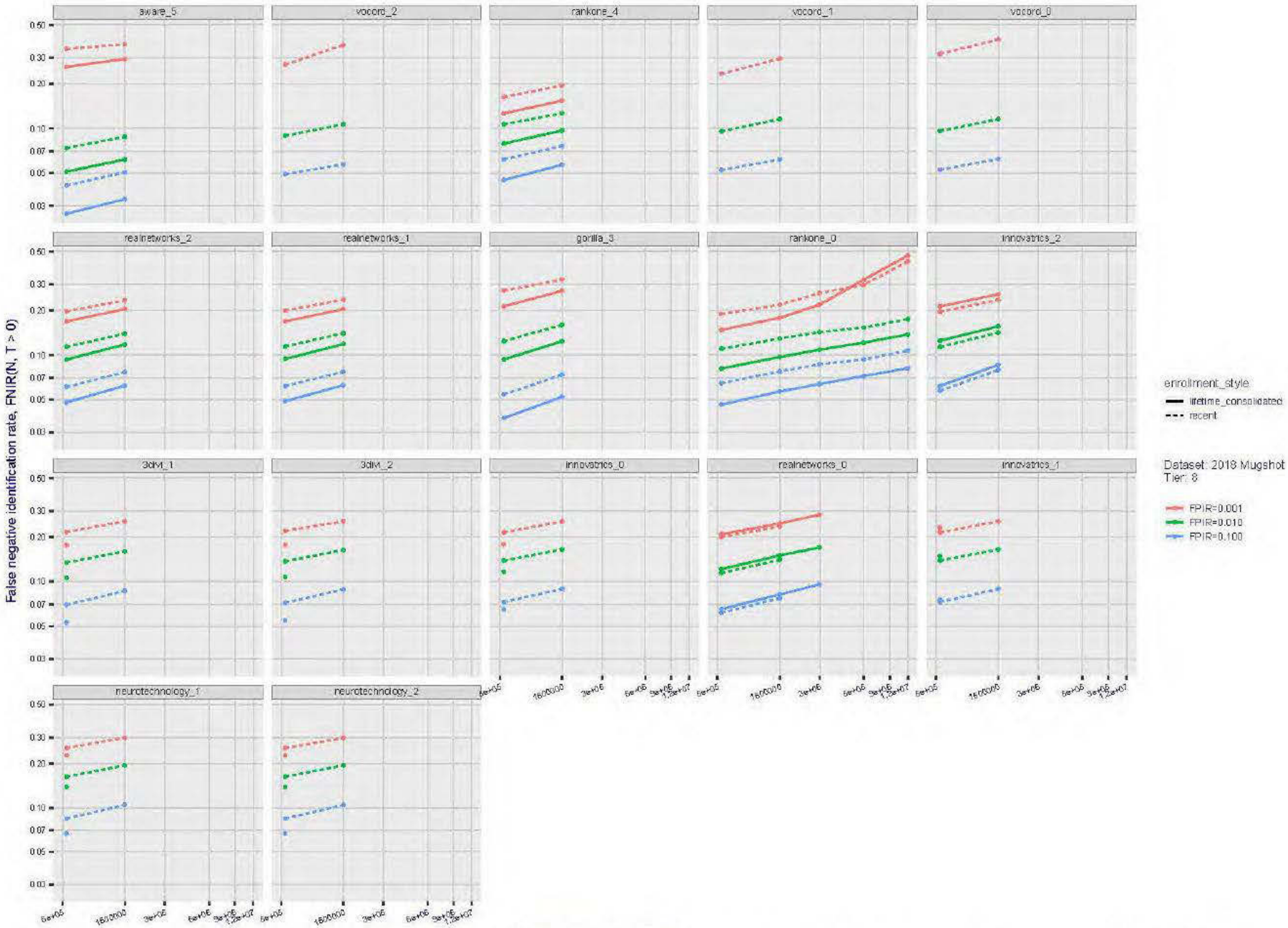


Figure 45: **[FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects.** *The figure shows FNIR(N, T) across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 5. Less accurate algorithms were not run on large N, so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by FNIR(N_b, 1, 0), then sorting by median FNIR(N_b, T), N_b = 640 000.*

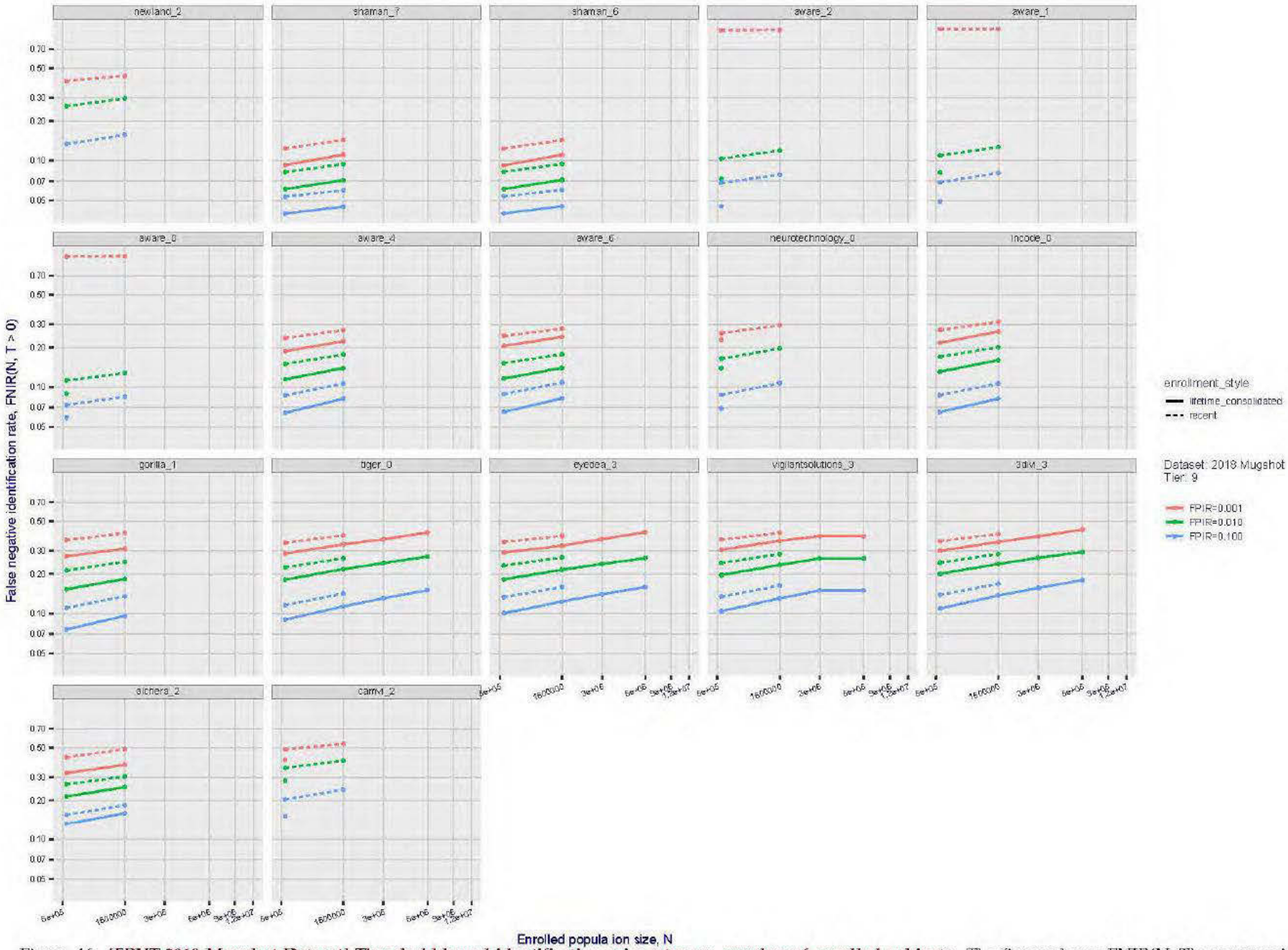


Figure 46: **[FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects.** *The figure shows FNIR(N, T) across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 5. Less accurate algorithms were not run on large N, so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by FNIR(N_b, 1, 0), then sorting by median FNIR(N_b, T), N_b = 640 000.*

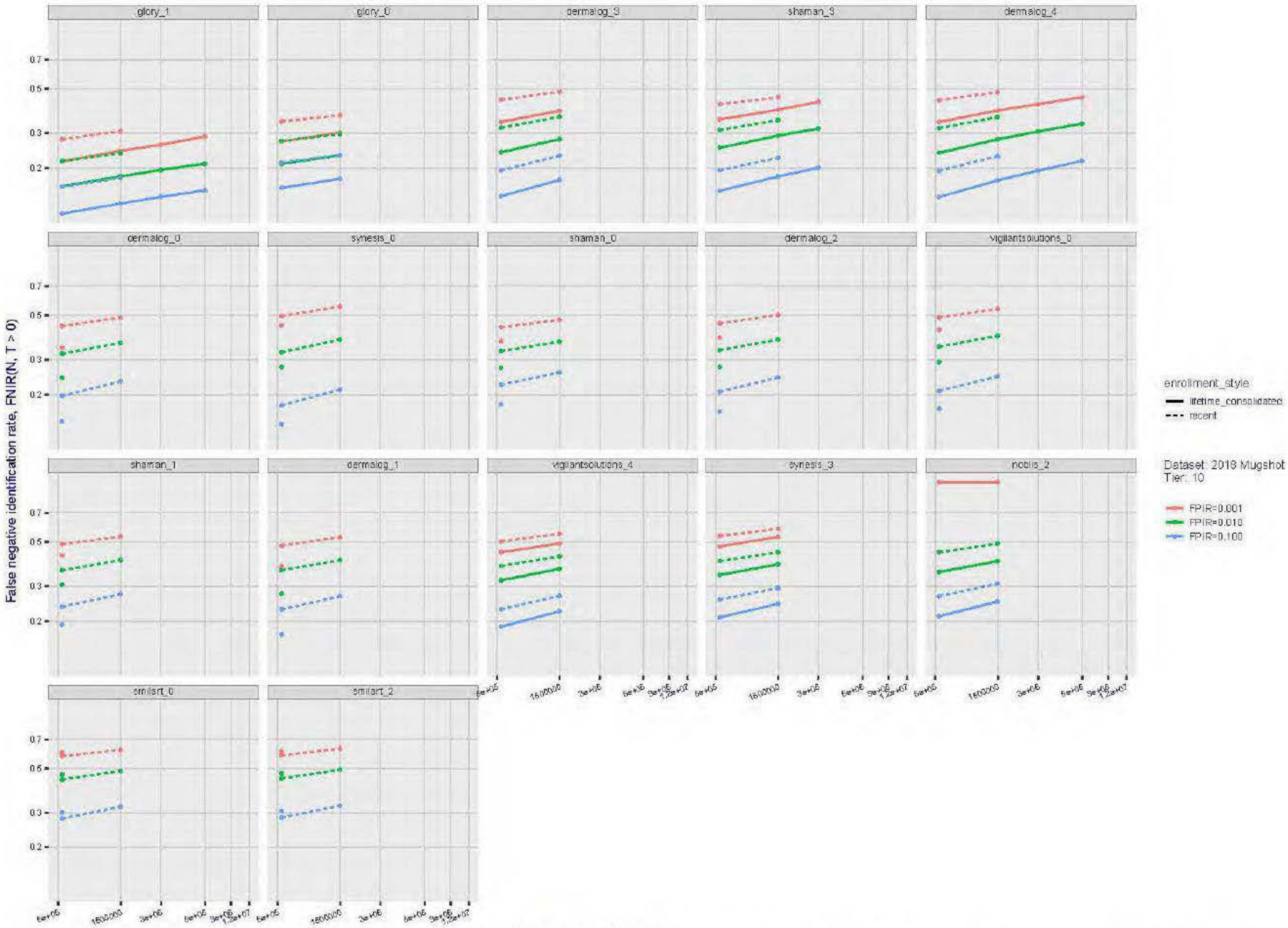


Figure 47: **[FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects.** *The figure shows FNIR(N, T) across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 5. Less accurate algorithms were not run on large N, so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by FNIR(N_b, 1, 0), then sorting by median FNIR(N_b, T), N_b = 640 000.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

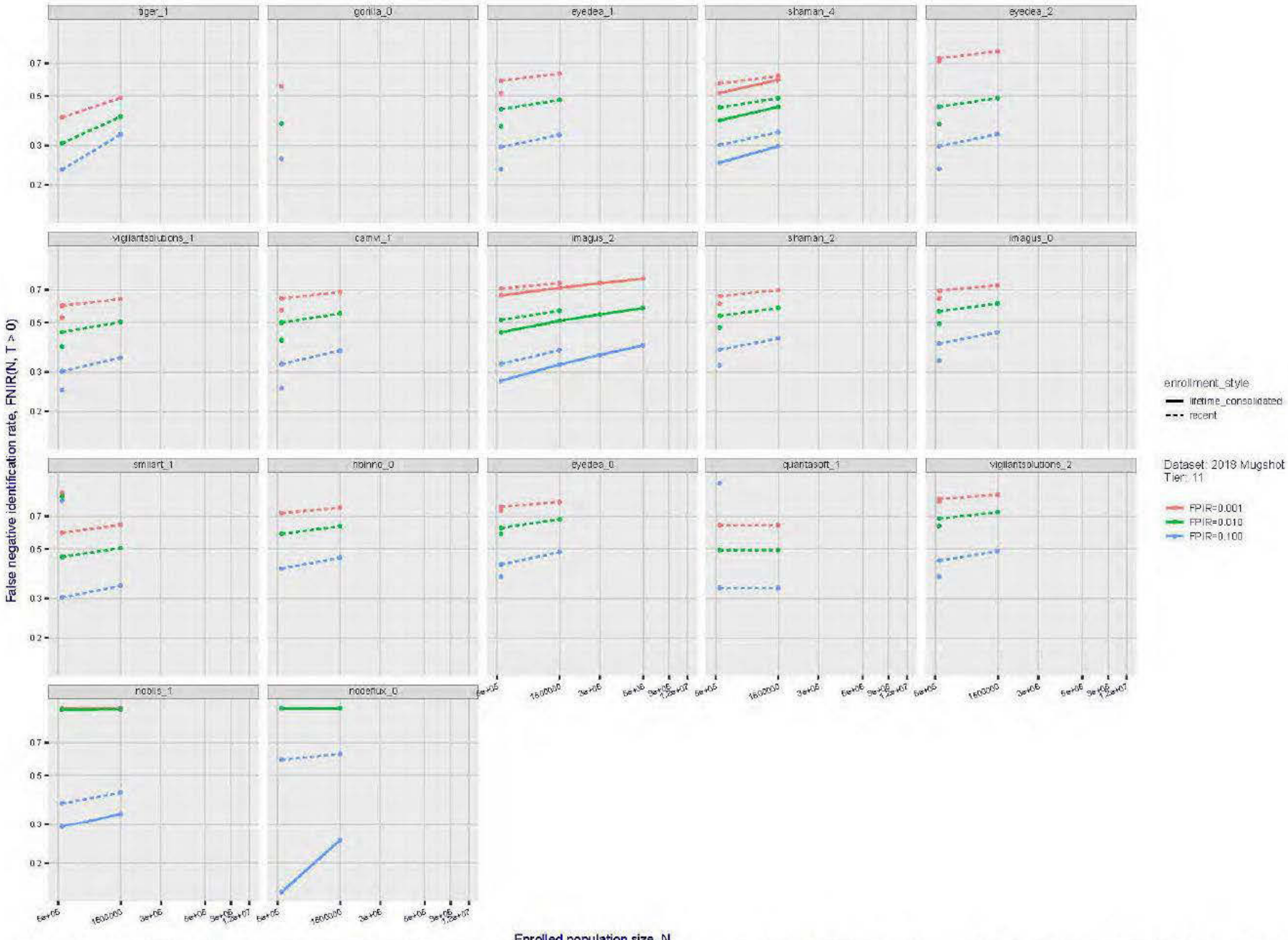


Figure 48: **[FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects.** *The figure shows FNIR(N, T) across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 5. Less accurate algorithms were not run on large N, so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by FNIR(N_b, 1, 0), then sorting by median FNIR(N_b, T), N_b = 640 000.*

FNIR(N, R, T) = False neg. identification rate N = Num. enrolled subjects T = Threshold T = 0 → Investigation
FPIR(N, T) = False pos. identification rate R = Num. candidates examined T > 0 → Identification

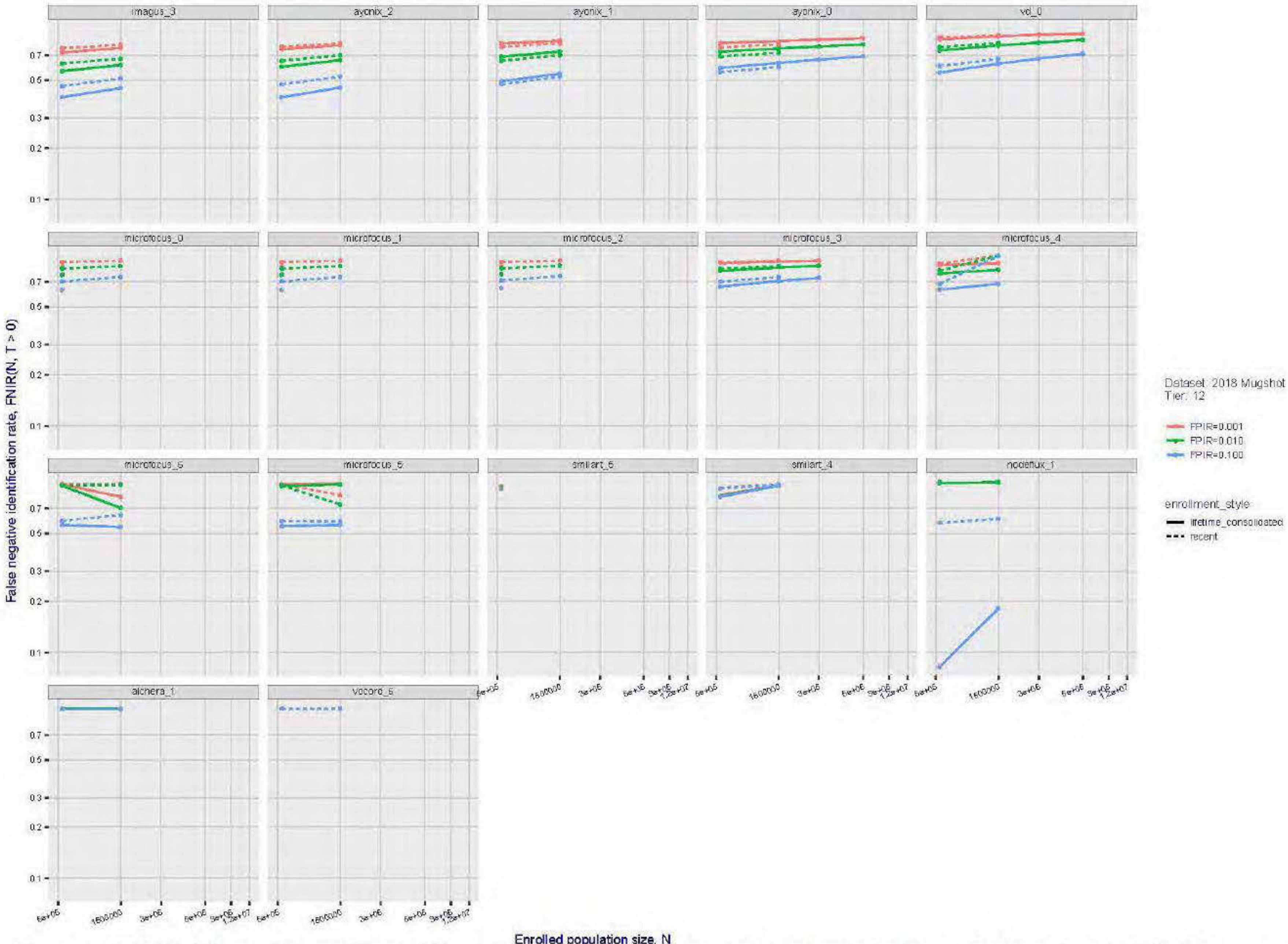


Figure 49: **[FRVT-2018 Mugshot Dataset] Threshold-based identification miss rates vs. number of enrolled subjects.** *The figure shows FNIR(N, T) across various gallery sizes when the threshold is set to achieve the given FPIRs. The rank criterion is irrelevant at high thresholds as mates are always at rank 1. The results are computed from the trials listed in rows 1-10 of Table 5. Less accurate algorithms were not run on large N, so results are missing. For clarity, results are sorted and reported into tiers spanning multiple pages. The tiering criteria is complicated: First paging by FNIR(N_b, 1, 0), then sorting by median FNIR(N_b, T), N_b = 640 000.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

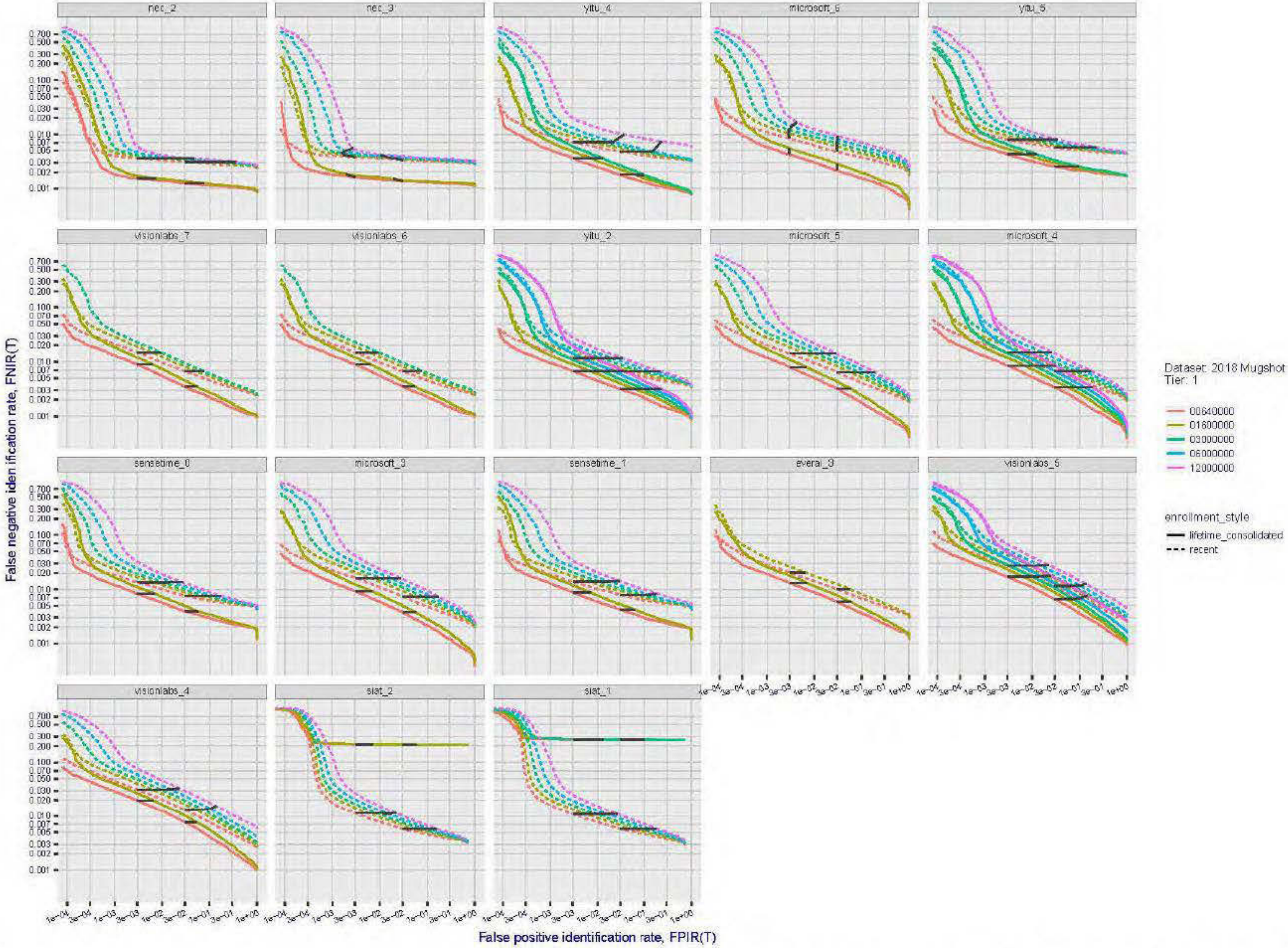


Figure 50: **[FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates.** *The figure shows miss rates FNIR(N, L, T) as a function of FPIR(N, T), with N ranging from 640 000 to 12 000 000 as noted in rows 1-10 of Table 5. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, FPIR(T) rises with N, and mate scores are independent of N. Other algorithms adjust scores in an attempt to make FPIR independent of N.*

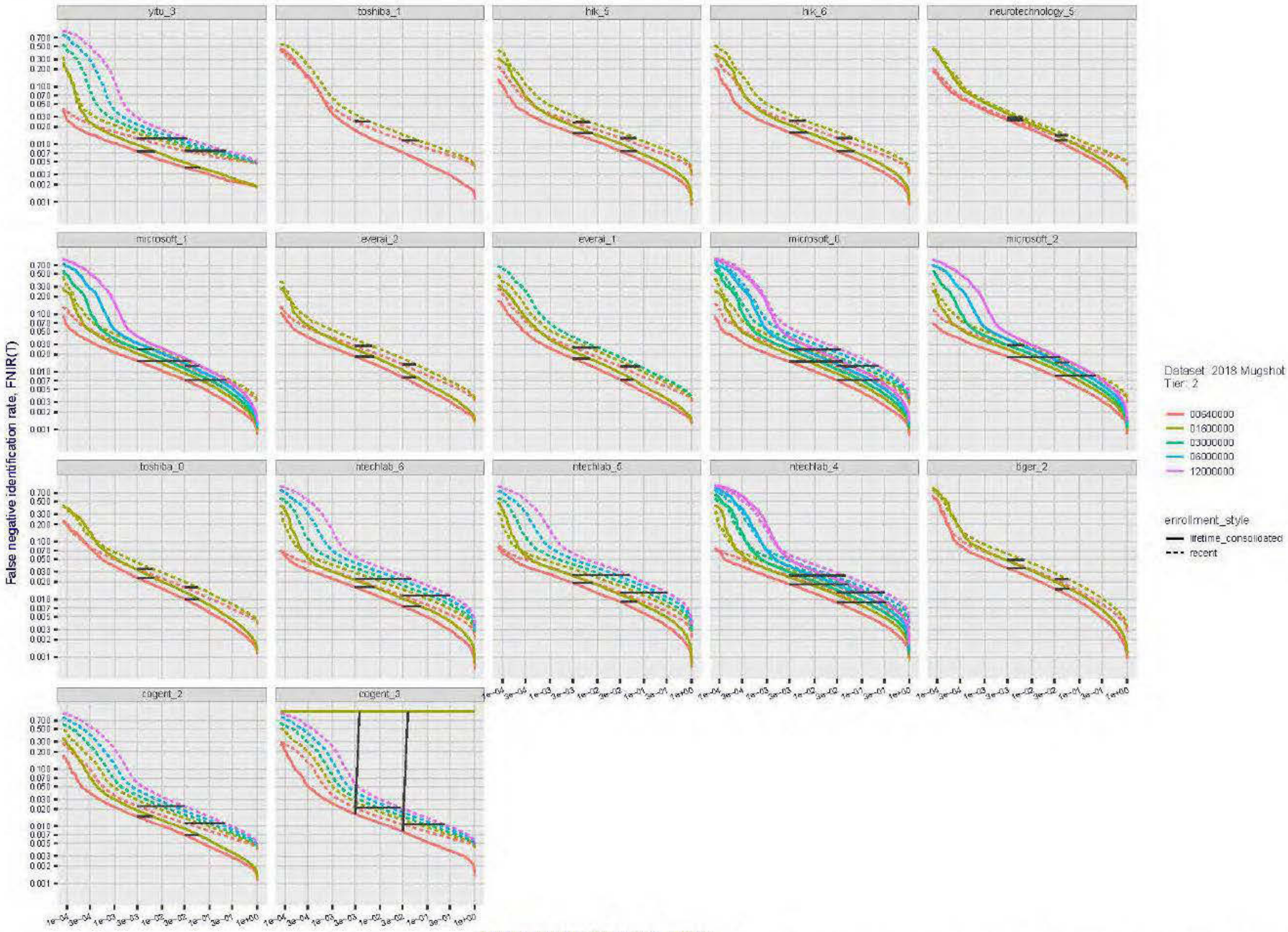


Figure 51: **[FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates.** *The figure shows miss rates FNIR(N, L, T) as a function of FPIR(N, T), with N ranging from 640 000 to 12 000 000 as noted in rows 1-10 of Table 5. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, FPIR(T) rises with N, and mate scores are independent of N. Other algorithms adjust scores in an attempt to make FPIR independent of N.*

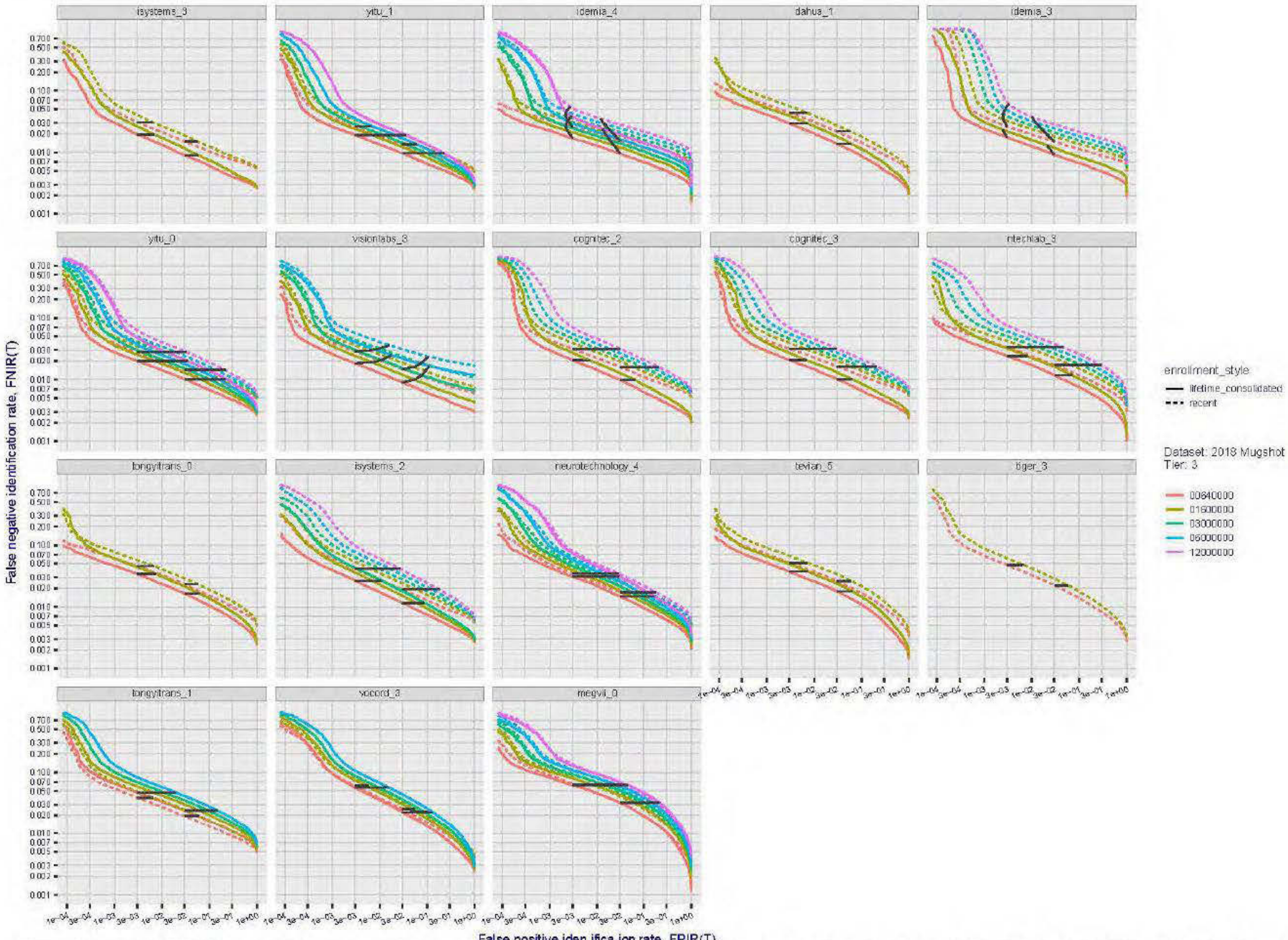


Figure 52: **[FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates.** *The figure shows miss rates FNIR(N, L, T) as a function of FPIR(N, T), with N ranging from 640 000 to 12 000 000 as noted in rows 1-10 of Table 5. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, FPIR(T) rises with N, and mate scores are independent of N. Other algorithms adjust scores in an attempt to make FPIR independent of N.*

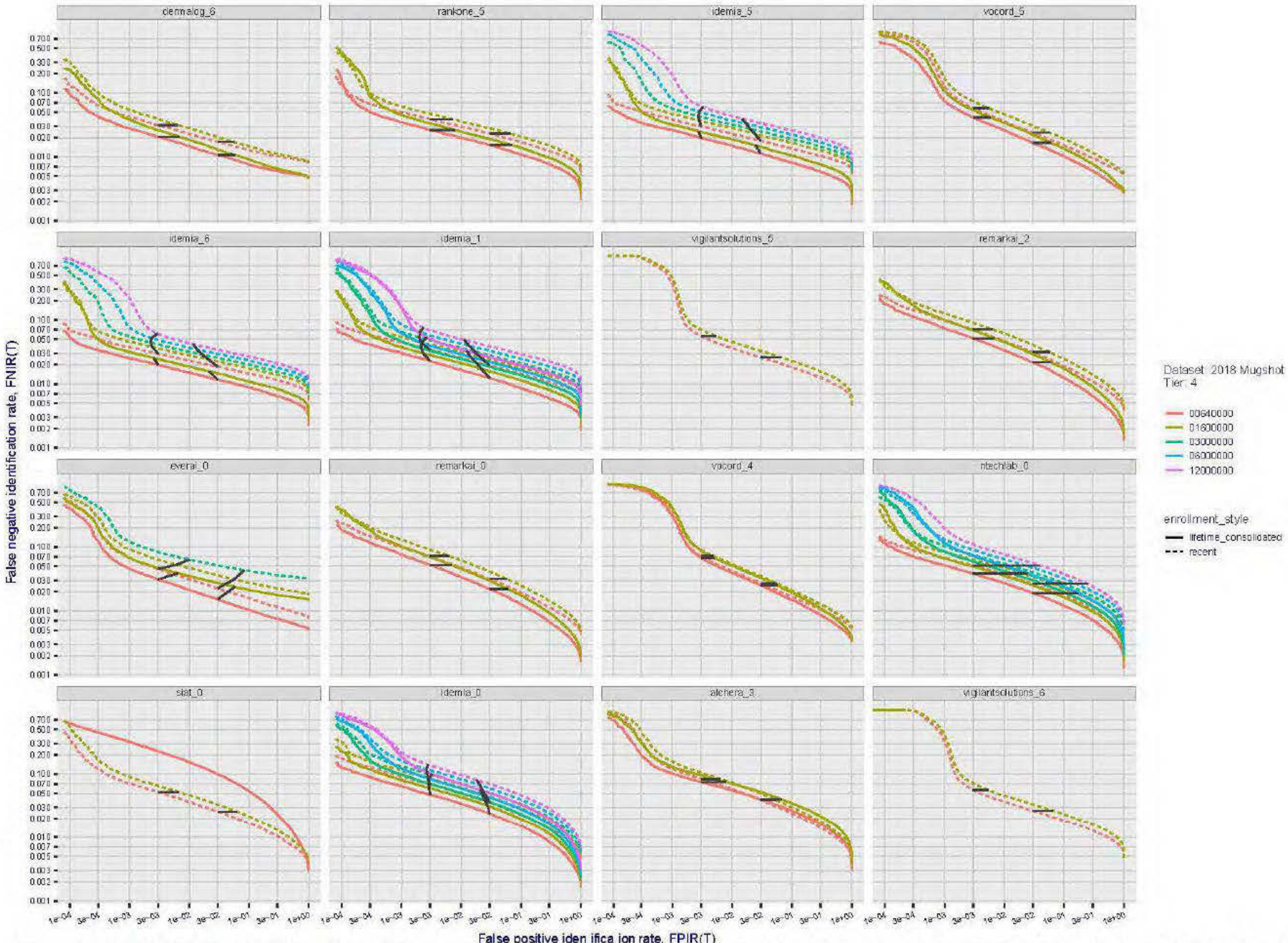


Figure 53: **[FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates.** *The figure shows miss rates FNIR(N, L, T) as a function of FPIR(N, T), with N ranging from 640 000 to 12 000 000 as noted in rows 1-10 of Table 5. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, FPIR(T) rises with N, and mate scores are independent of N. Other algorithms adjust scores in an attempt to make FPIR independent of N.*

FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

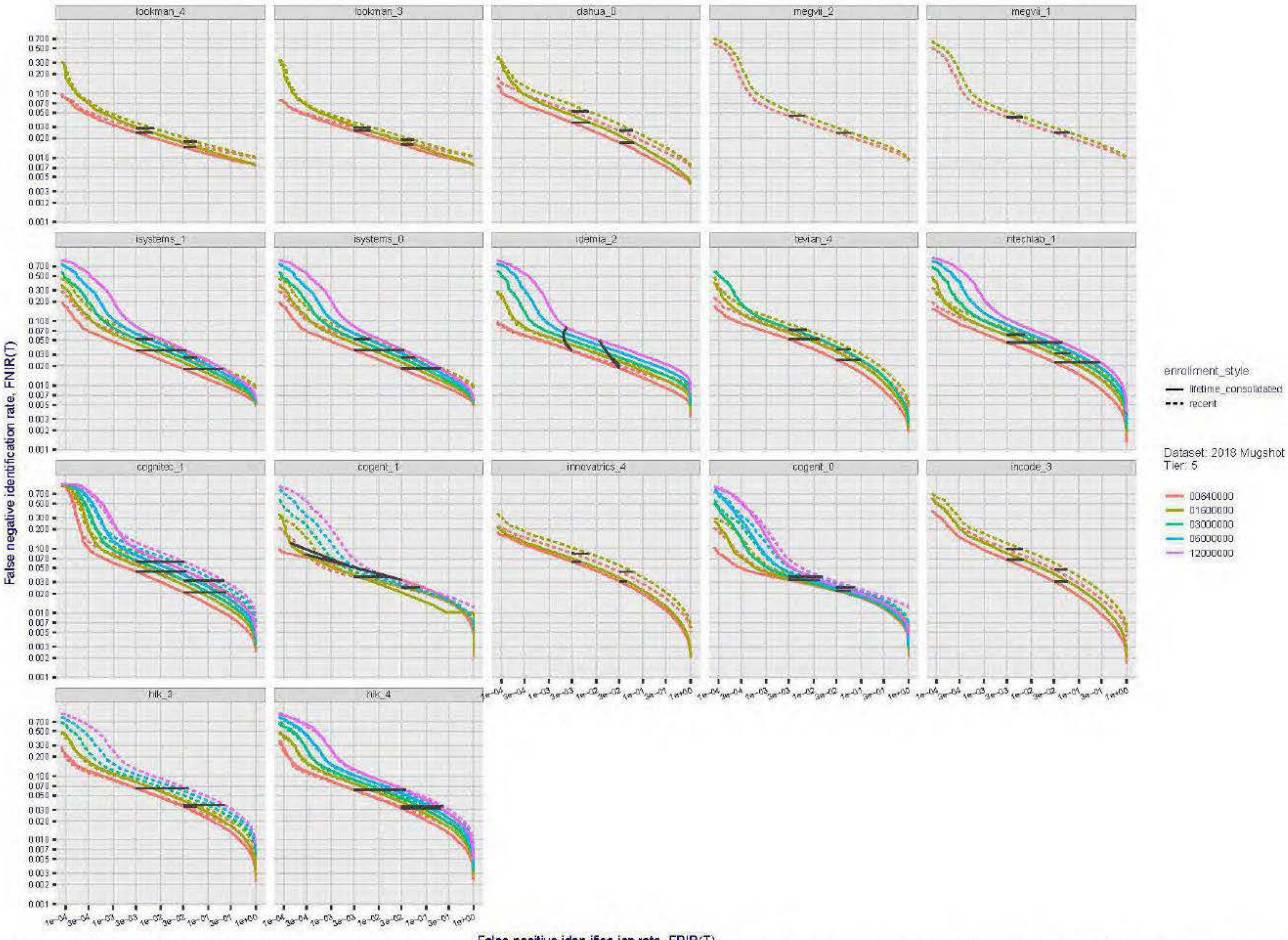


Figure 54: **[FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates.** *The figure shows miss rates FNIR(N, L, T) as a function of FPIR(N, T), with N ranging from 640 000 to 12 000 000 as noted in rows 1-10 of Table 5. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, FPIR(T) rises with N, and mate scores are independent of N. Other algorithms adjust scores in an attempt to make FPIR independent of N.*

FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

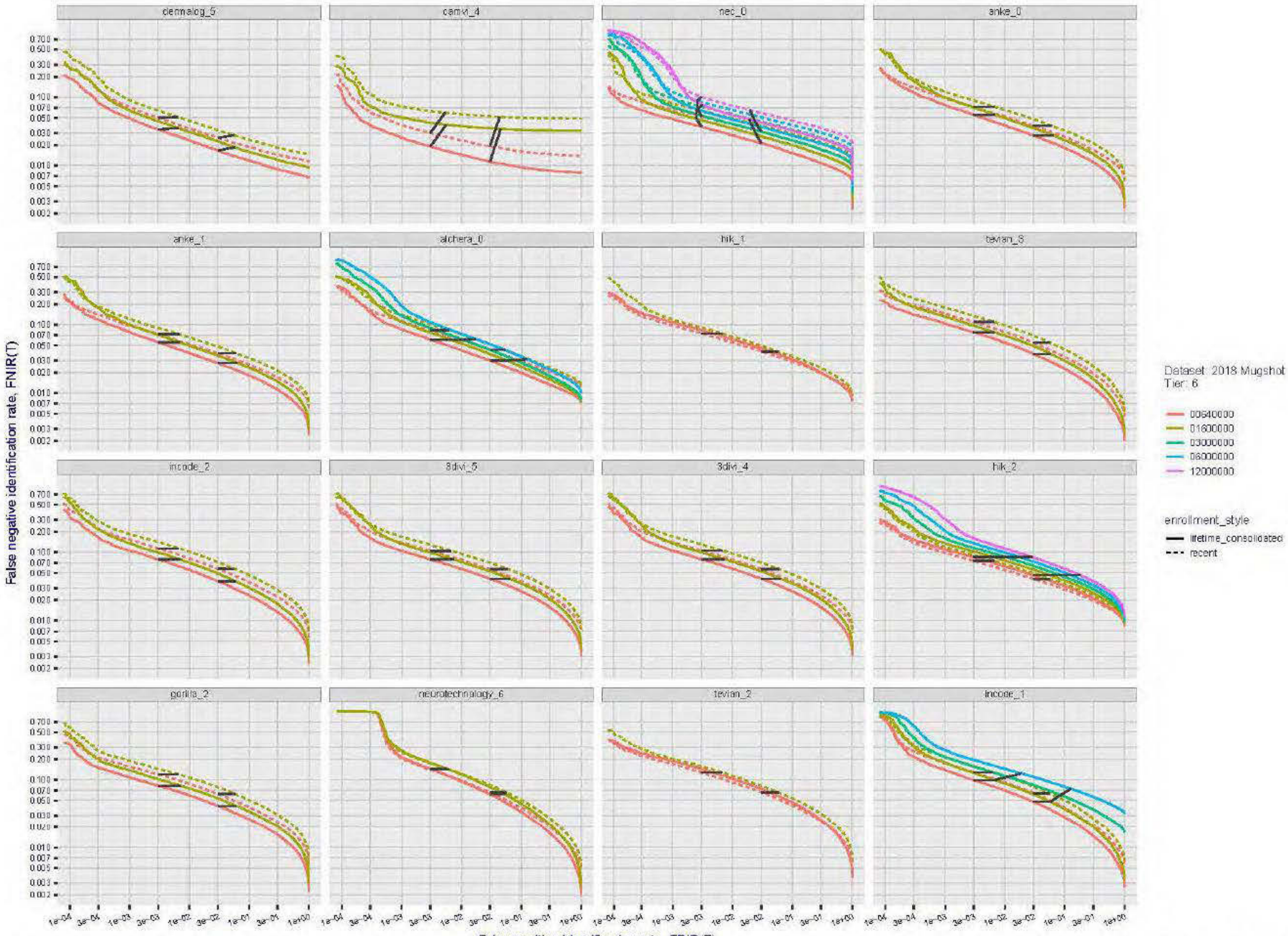


Figure 55: **[FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates.** *The figure shows miss rates FNIR(N, L, T) as a function of FPIR(N, T), with N ranging from 640 000 to 12 000 000 as noted in rows 1-10 of Table 5. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, FPIR(T) rises with N, and mate scores are independent of N. Other algorithms adjust scores in an attempt to make FPIR independent of N.*

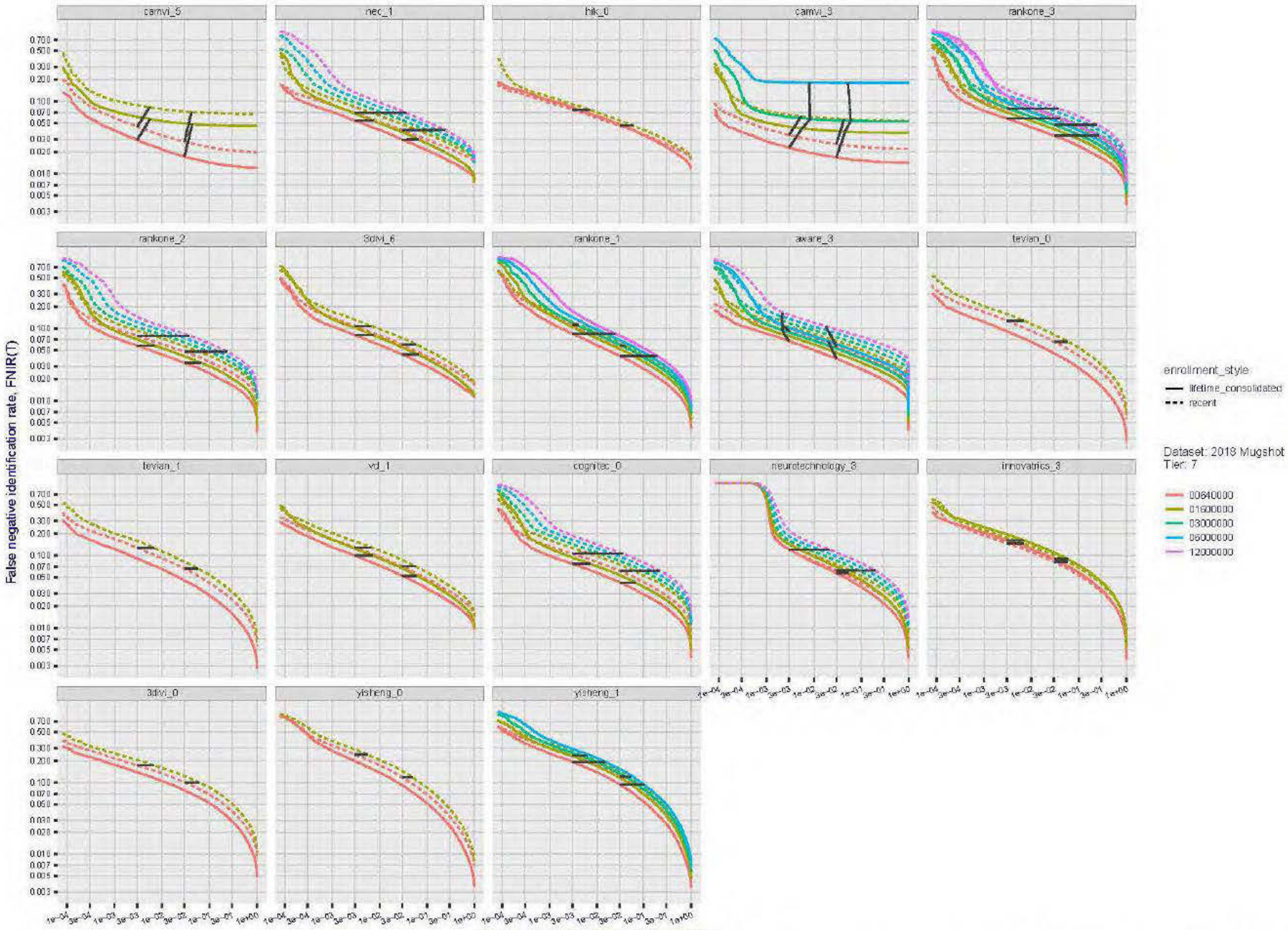


Figure 56: **[FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates.** *The figure shows miss rates FNIR(N, L, T) as a function of FPIR(N, T), with N ranging from 640 000 to 12 000 000 as noted in rows 1-10 of Table 5. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, FPIR(T) rises with N, and mate scores are independent of N. Other algorithms adjust scores in an attempt to make FPIR independent of N.*

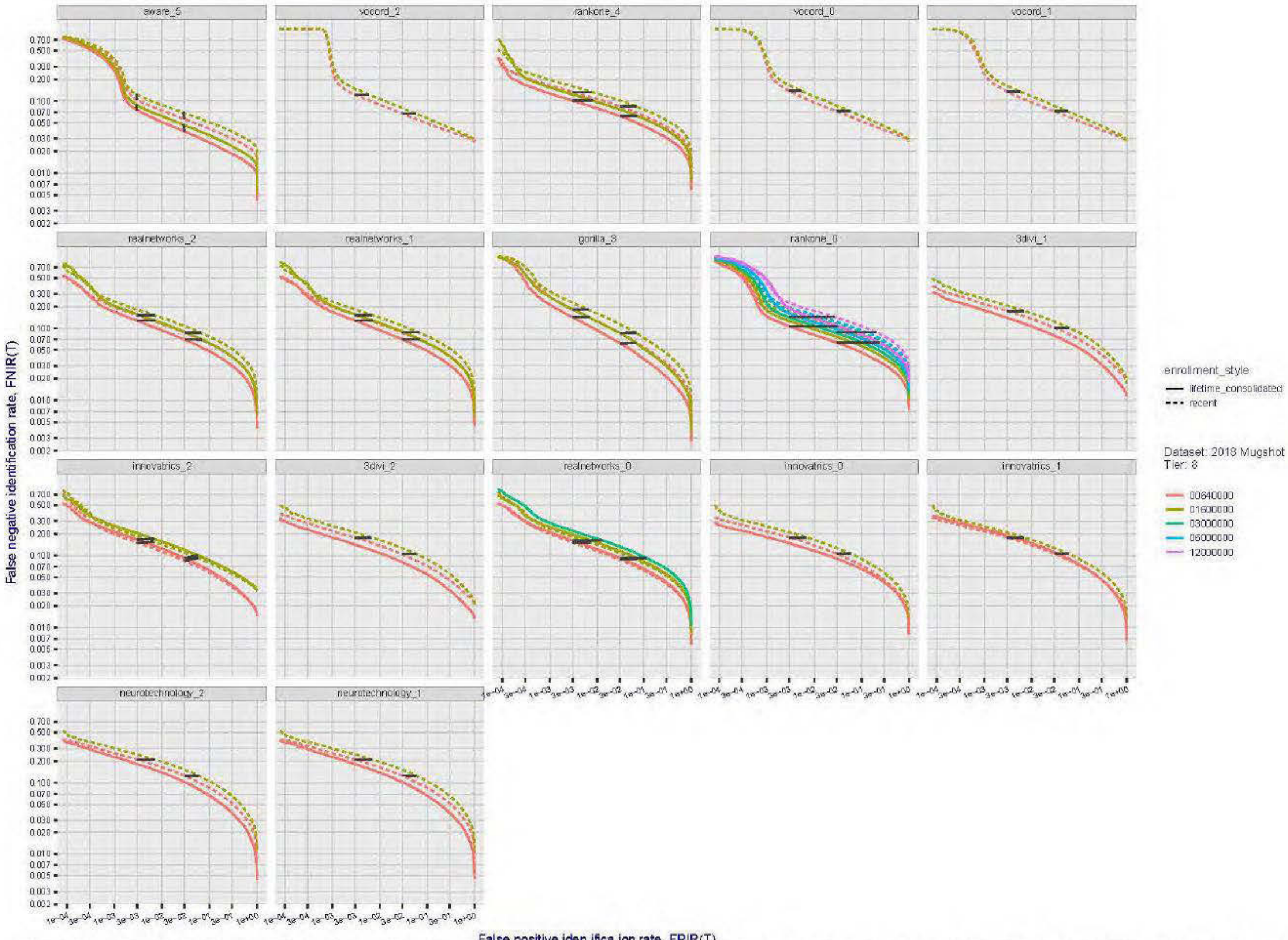


Figure 57: **[FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates.** *The figure shows miss rates FNIR(N, L, T) as a function of FPIR(N, T), with N ranging from 640 000 to 12 000 000 as noted in rows 1-10 of Table 5. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, FPIR(T) rises with N, and mate scores are independent of N. Other algorithms adjust scores in an attempt to make FPIR independent of N.*

2019/09/11 17:24:52

FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

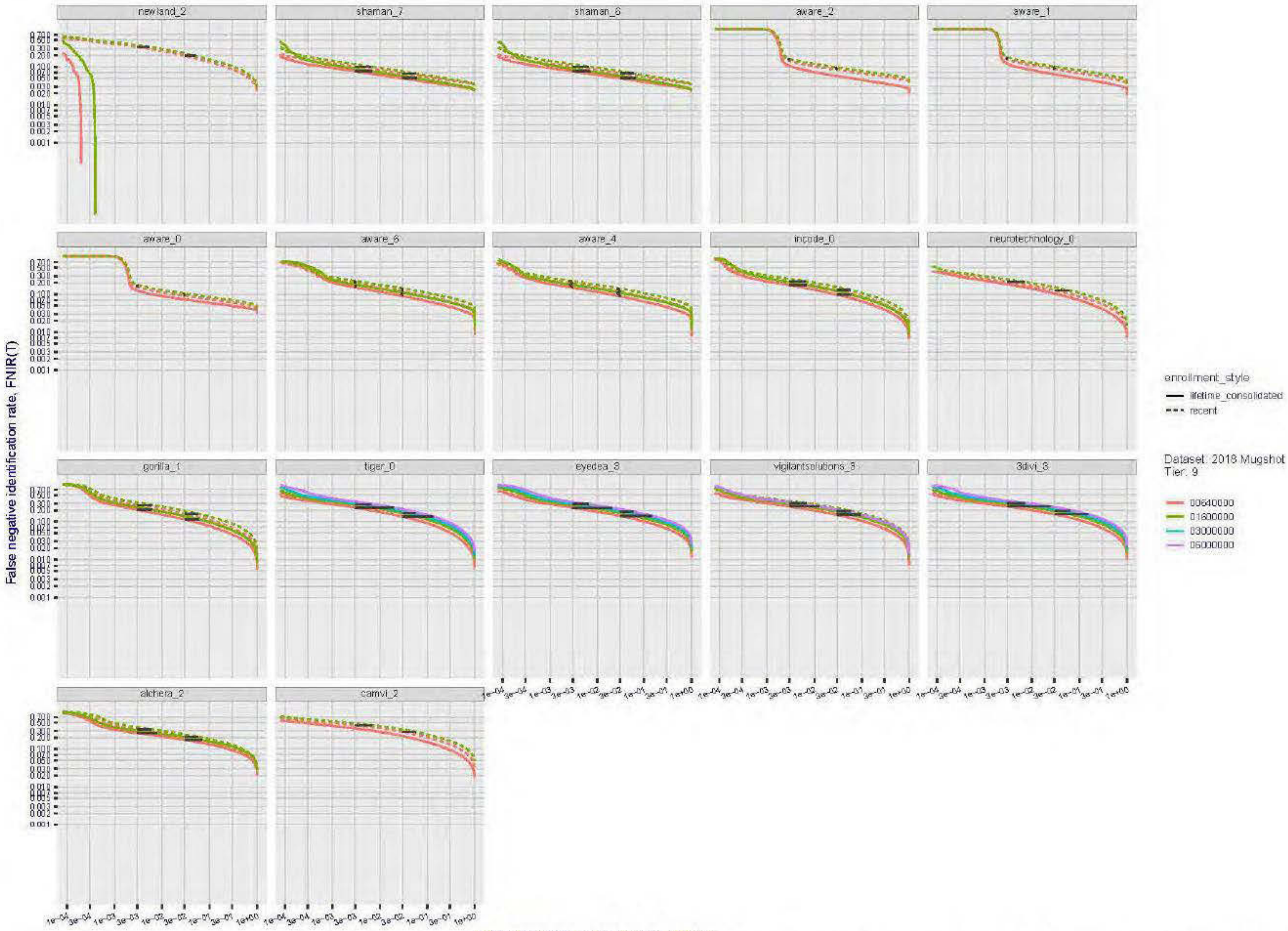


Figure 58: **[FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates.** *The figure shows miss rates FNIR(N, L, T) as a function of FPIR(N, T), with N ranging from 640 000 to 12 000 000 as noted in rows 1-10 of Table 5. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, FPIR(T) rises with N, and mate scores are independent of N. Other algorithms adjust scores in an attempt to make FPIR independent of N.*

FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

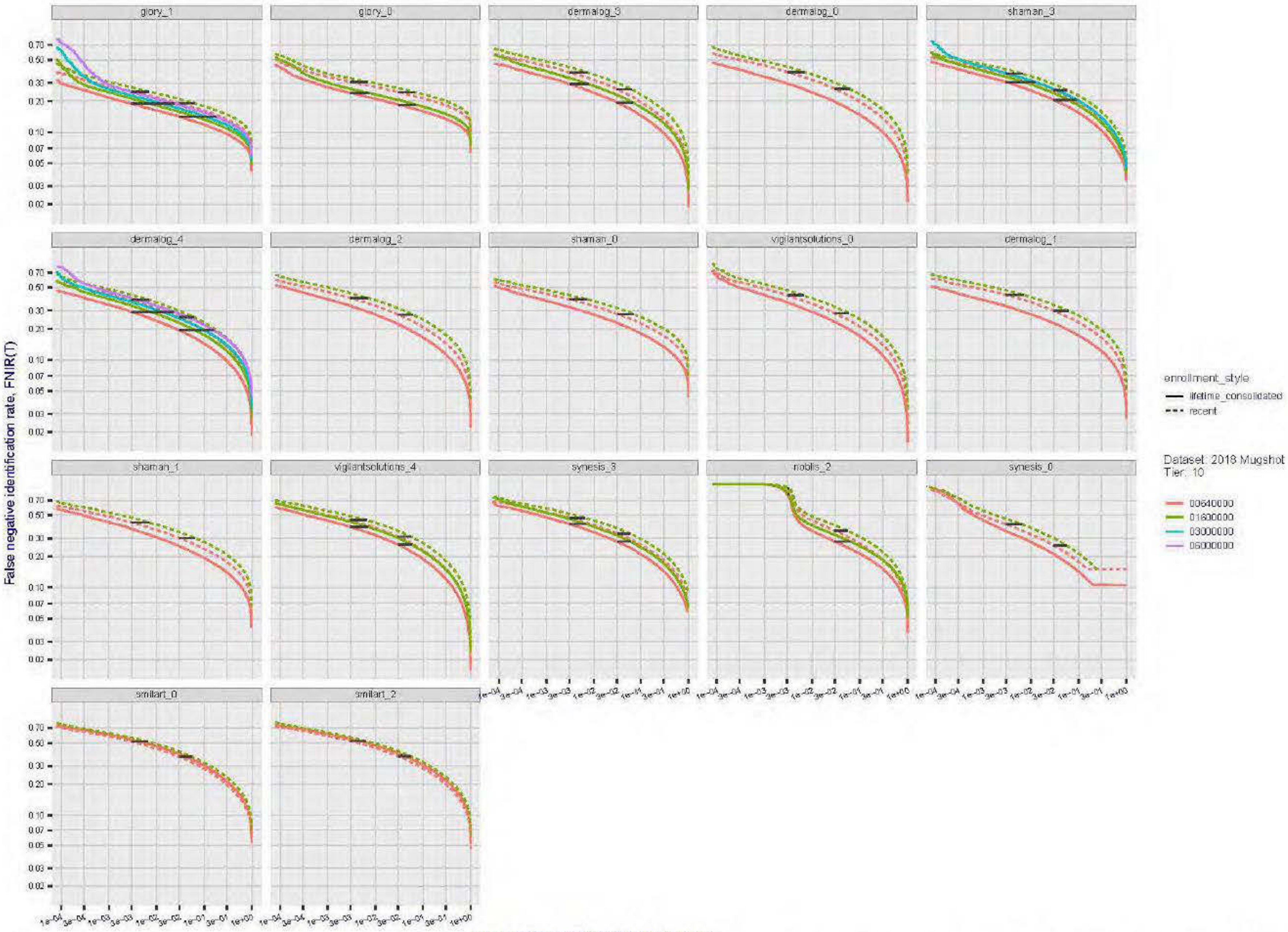


Figure 59: **[FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates.** *The figure shows miss rates FNIR(N, L, T) as a function of FPIR(N, T), with N ranging from 640 000 to 12 000 000 as noted in rows 1-10 of Table 5. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, FPIR(T) rises with N, and mate scores are independent of N. Other algorithms adjust scores in an attempt to make FPIR independent of N.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

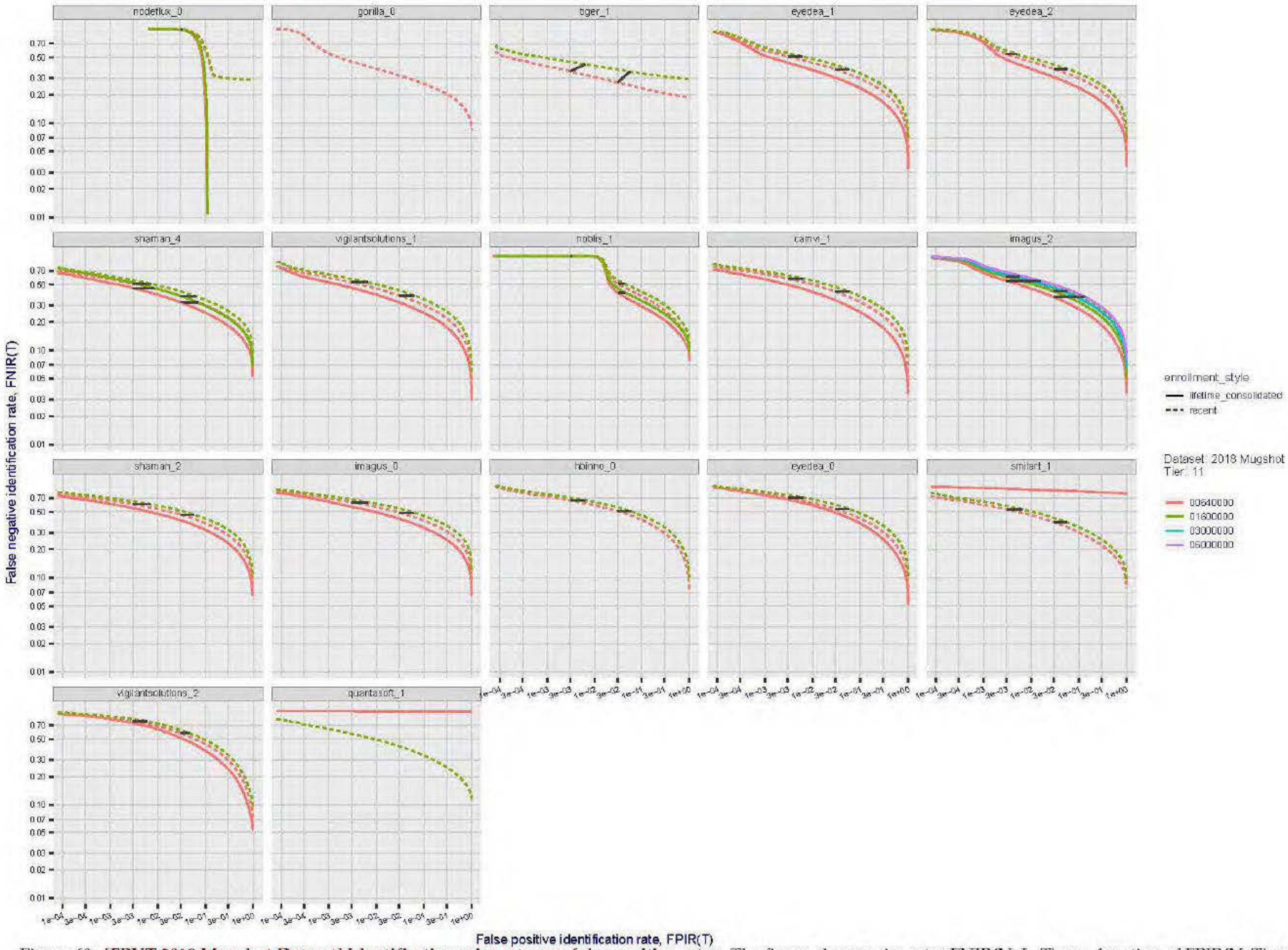


Figure 60: **[FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates.** *The figure shows miss rates FNIR(N, L, T) as a function of FPIR(N, T), with N ranging from 640 000 to 12 000 000 as noted in rows 1-10 of Table 5. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, FPIR(T) rises with N, and mate scores are independent of N. Other algorithms adjust scores in an attempt to make FPIR independent of N.*

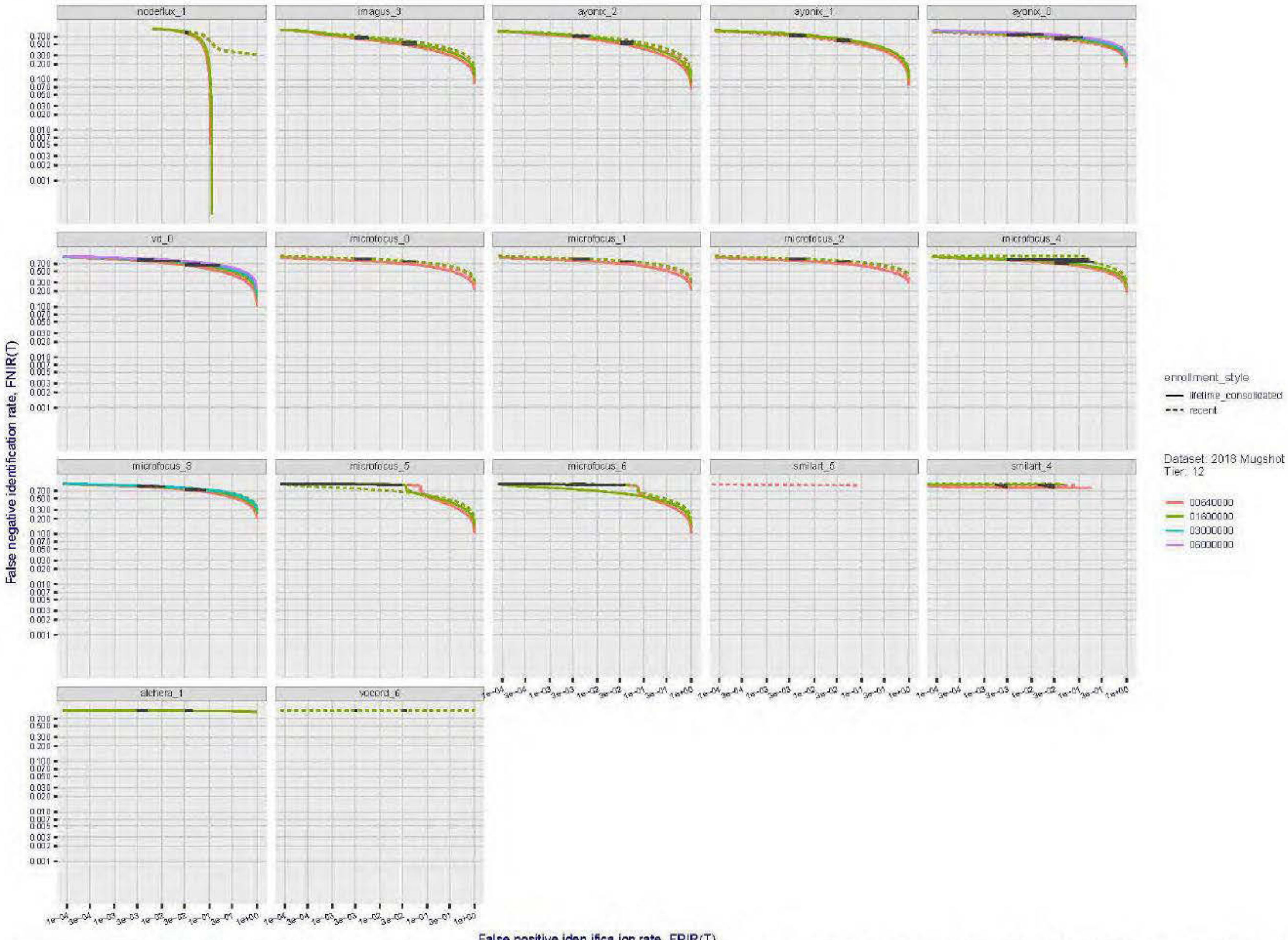


Figure 61: **[FRVT-2018 Mugshot Dataset] Identification miss rates vs. false positive rates.** *The figure shows miss rates FNIR(N, L, T) as a function of FPIR(N, T), with N ranging from 640 000 to 12 000 000 as noted in rows 1-10 of Table 5. These error tradeoff characteristics are useful for applications where a threshold must be elevated to limit false positives, such as when human reviewer labor is not matched to the volume of searches. Dark lines join points of equal threshold: If horizontal, FPIR(T) rises with N, and mate scores are independent of N. Other algorithms adjust scores in an attempt to make FPIR independent of N.*

Appendix B Effect of time-lapse: Accuracy after face ageing

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

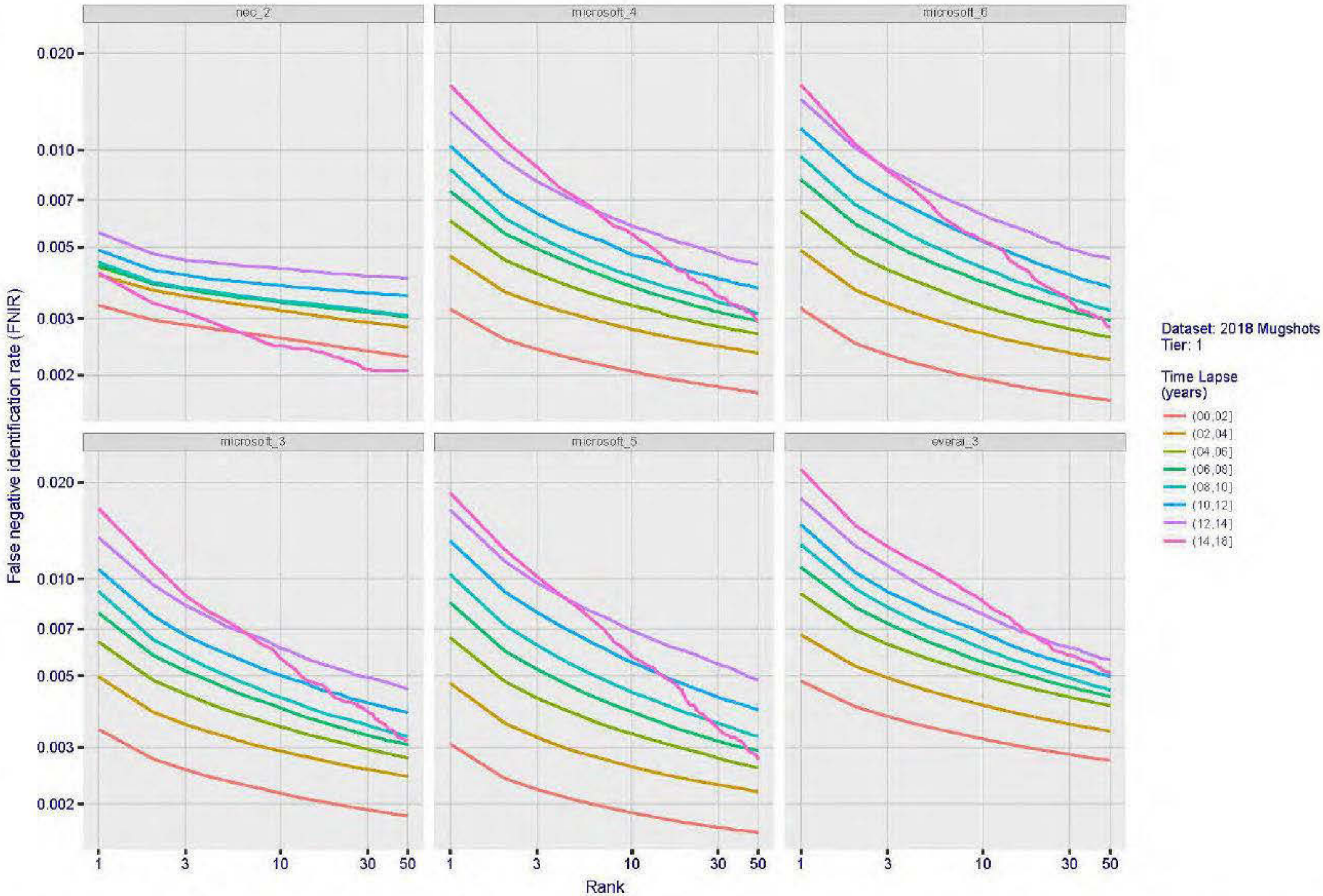


Figure 62: **[FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 5 and binned by number of years between search and initial enrollment.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

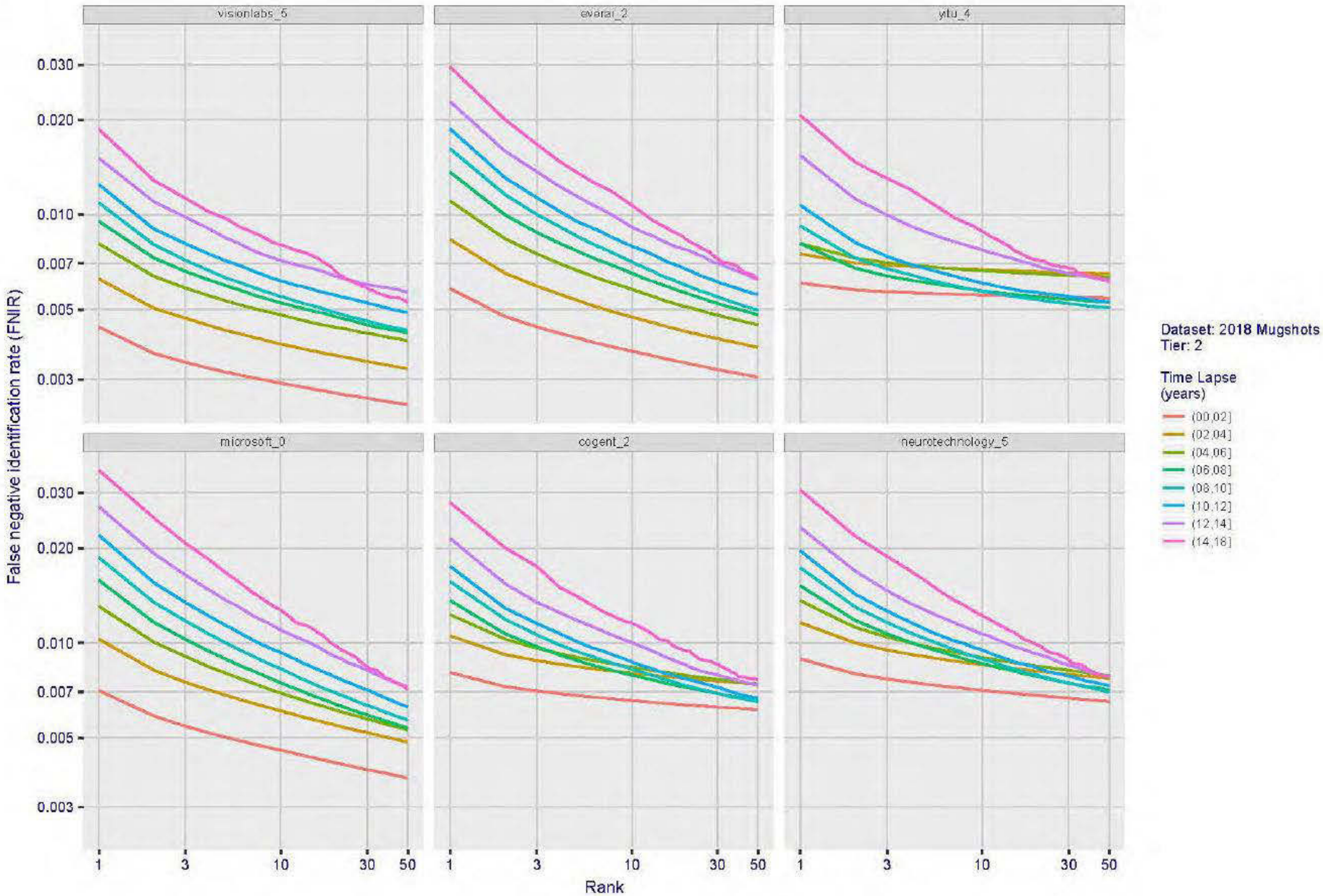


Figure 63: **[FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 5 and binned by number of years between search and initial enrollment.*

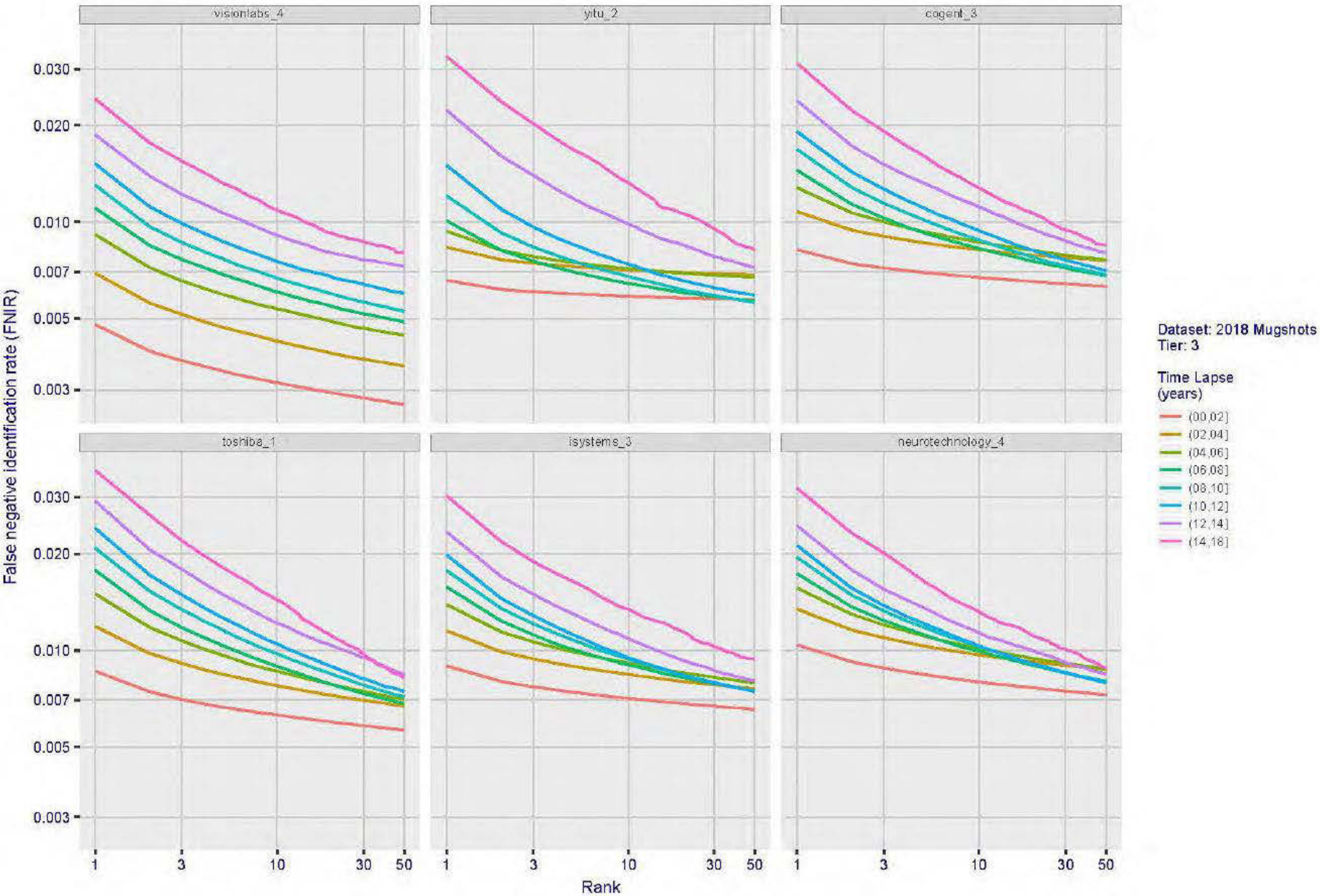


Figure 64: **[FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 5 and binned by number of years between search and initial enrollment.*

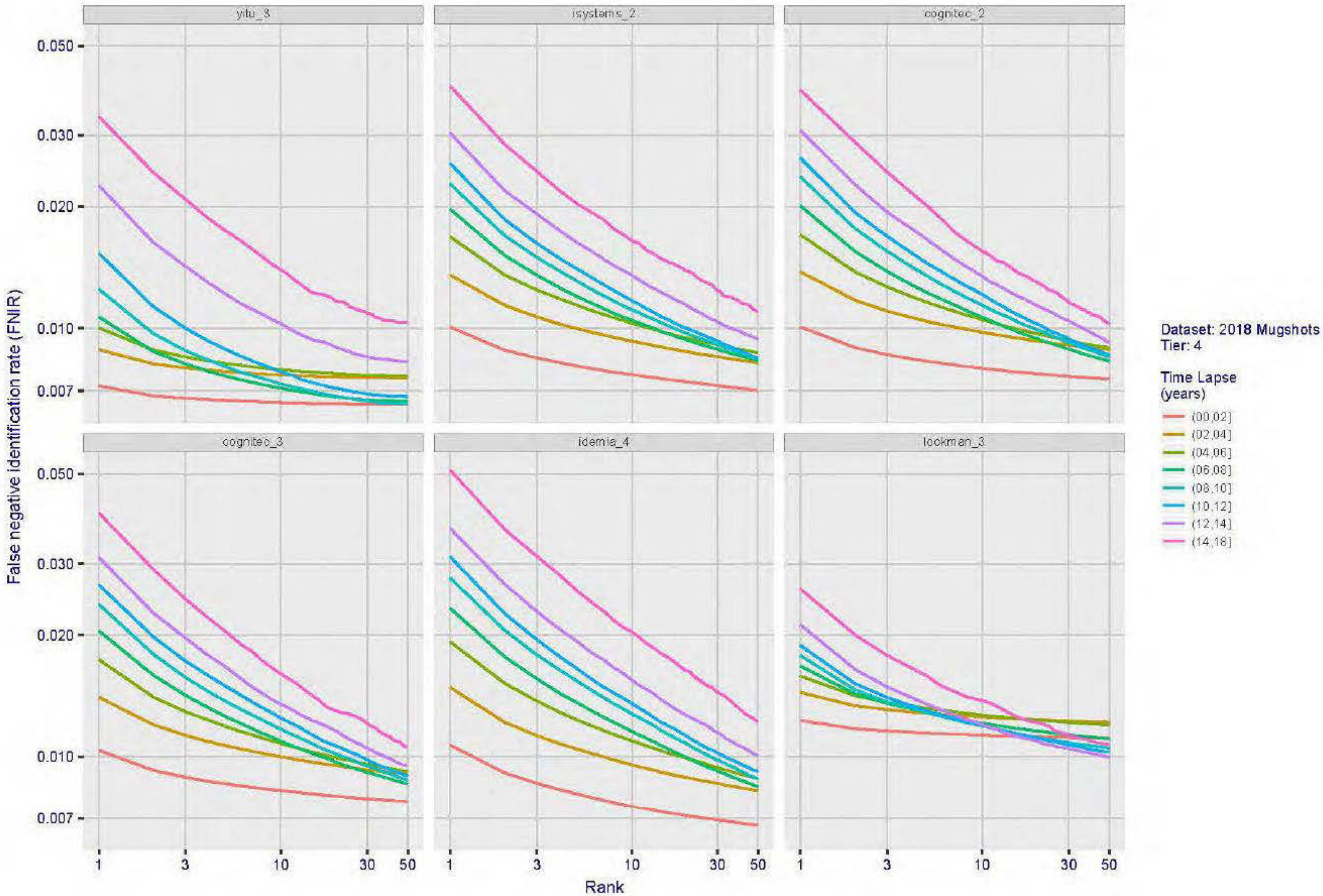


Figure 65: **[FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 5 and binned by number of years between search and initial enrollment.*

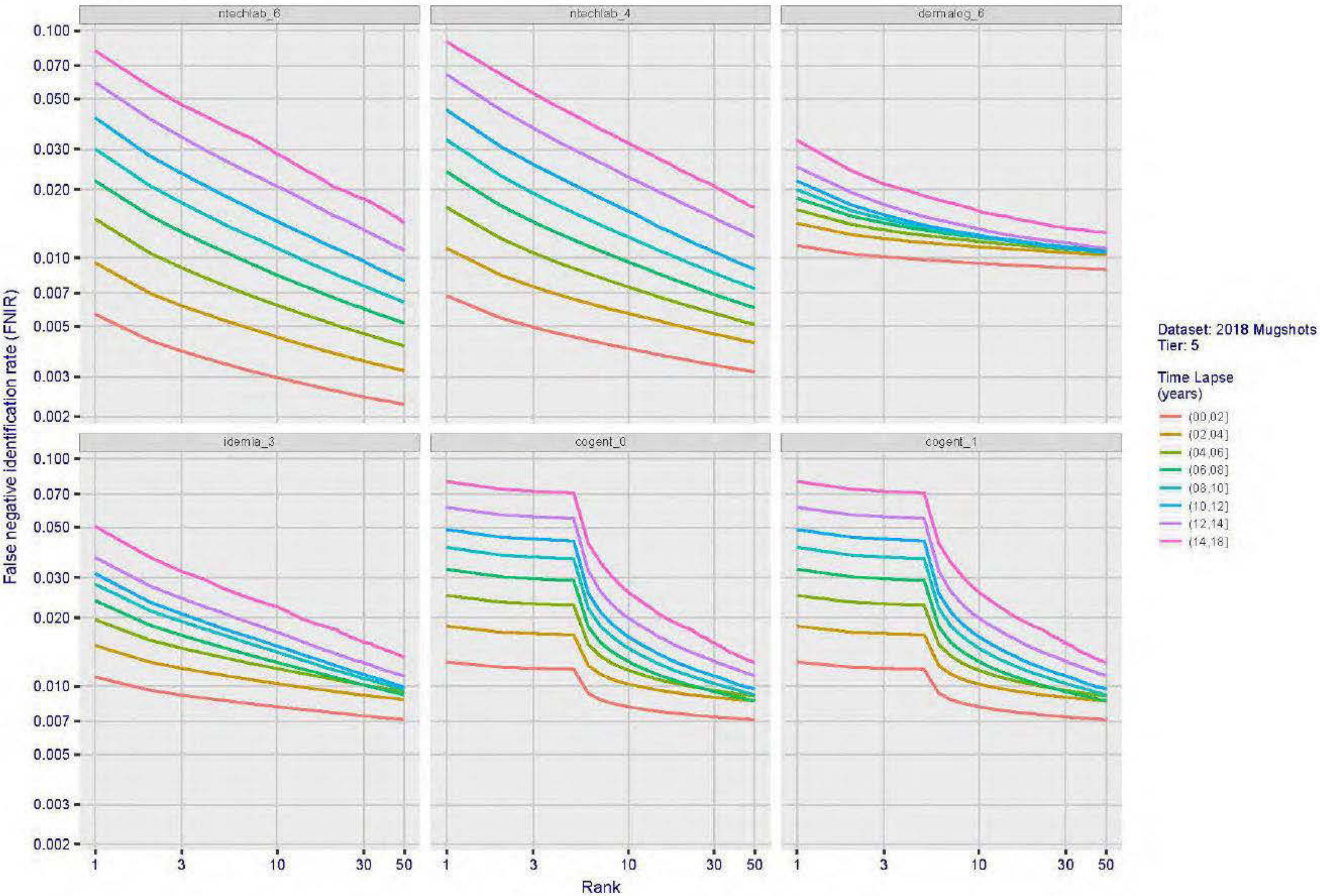


Figure 66: **[FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 5 and binned by number of years between search and initial enrollment.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

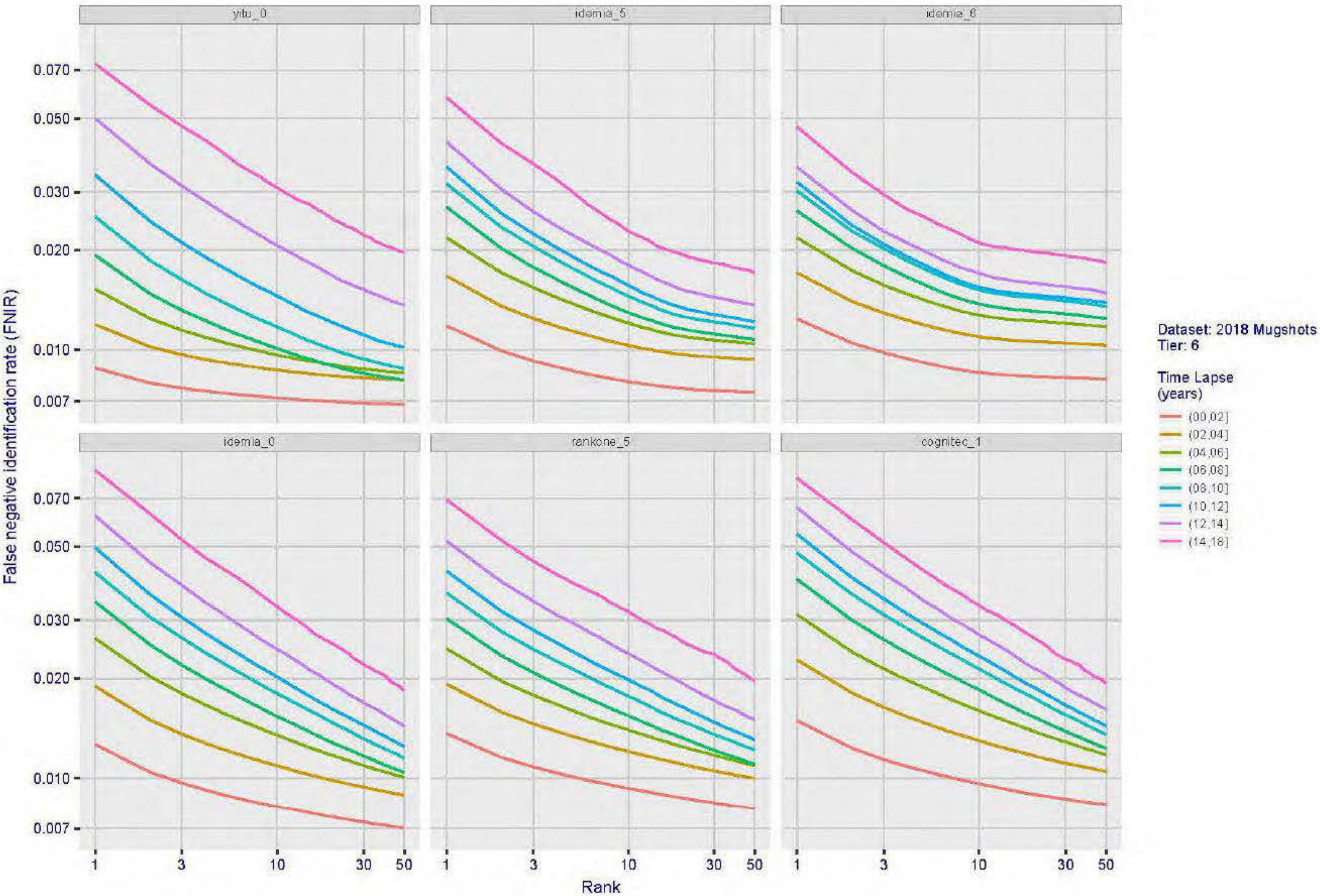


Figure 67: **[FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 5 and binned by number of years between search and initial enrollment.*

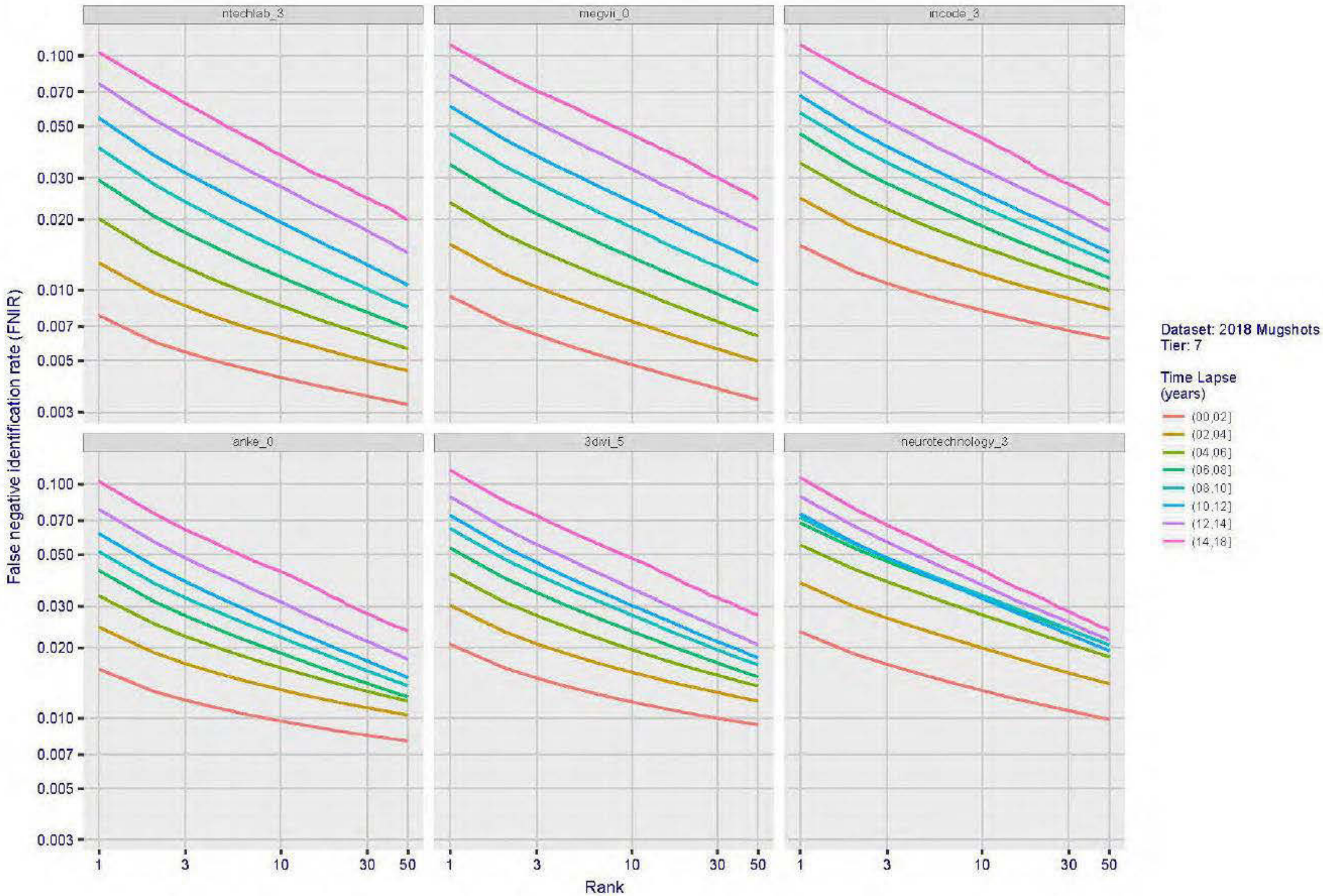


Figure 68: **[FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 5 and binned by number of years between search and initial enrollment.*

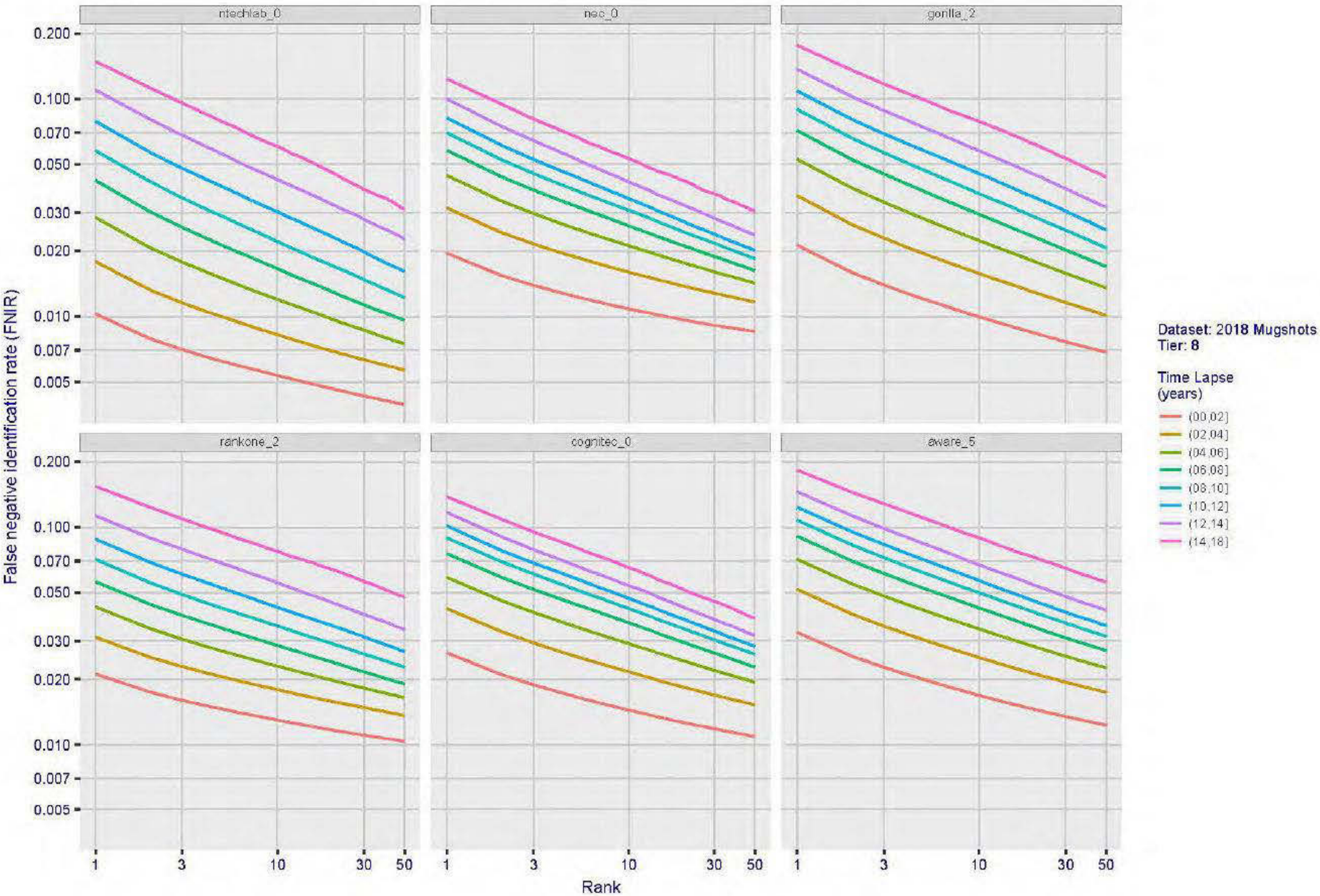


Figure 69: **[FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 5 and binned by number of years between search and initial enrollment.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

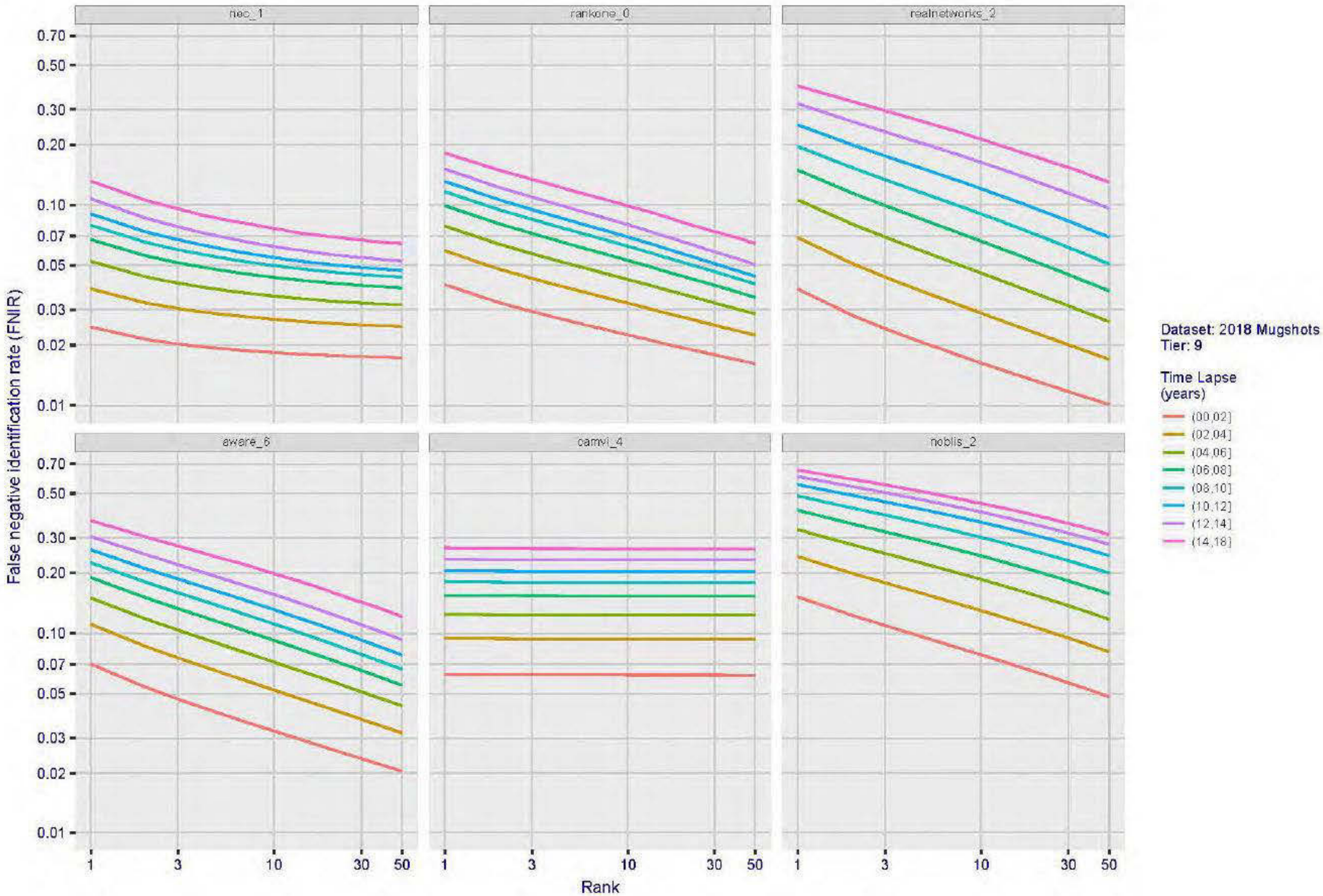


Figure 70: **[FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 5 and binned by number of years between search and initial enrollment.*

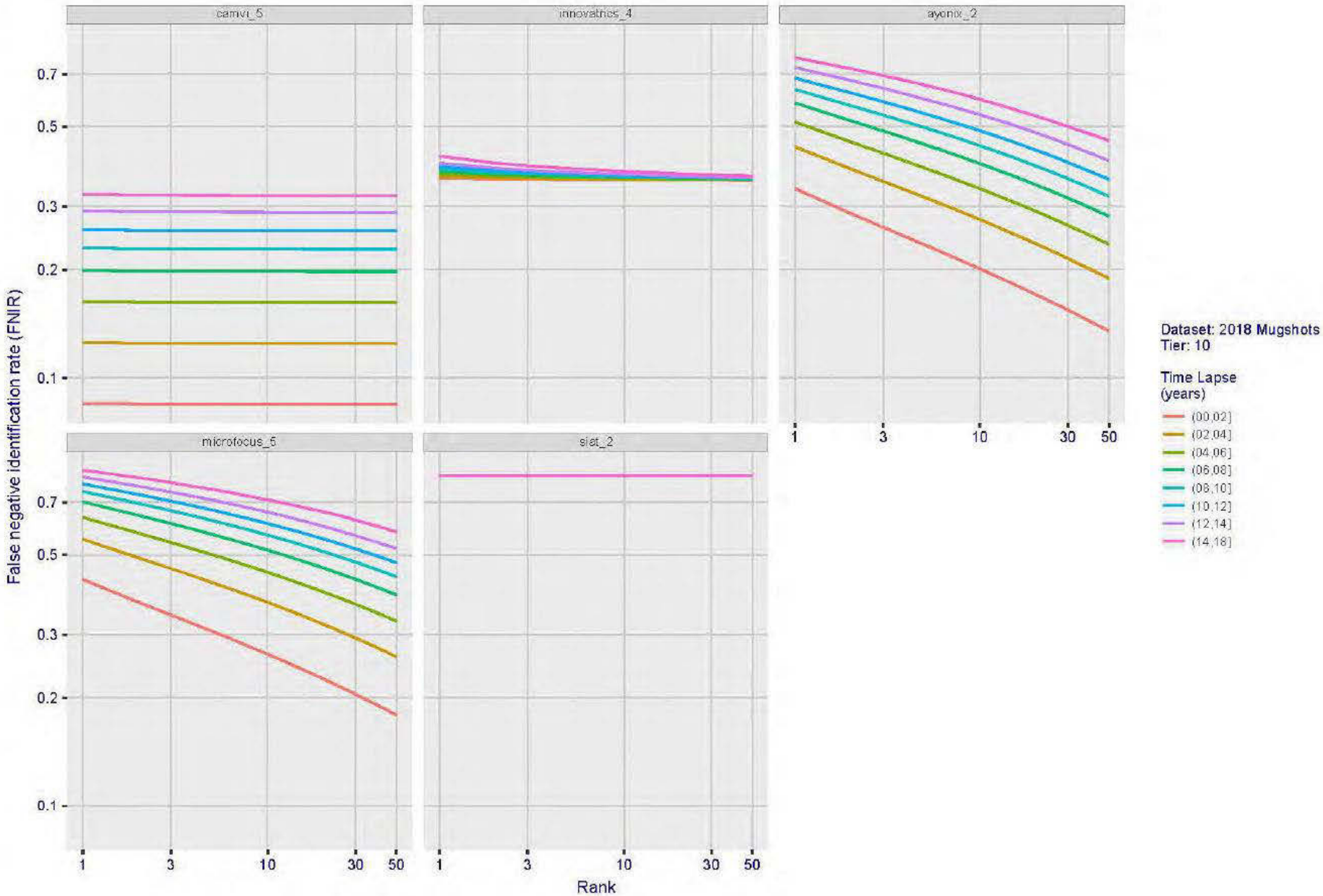


Figure 71: **[FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. rank by time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 5 and binned by number of years between search and initial enrollment.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

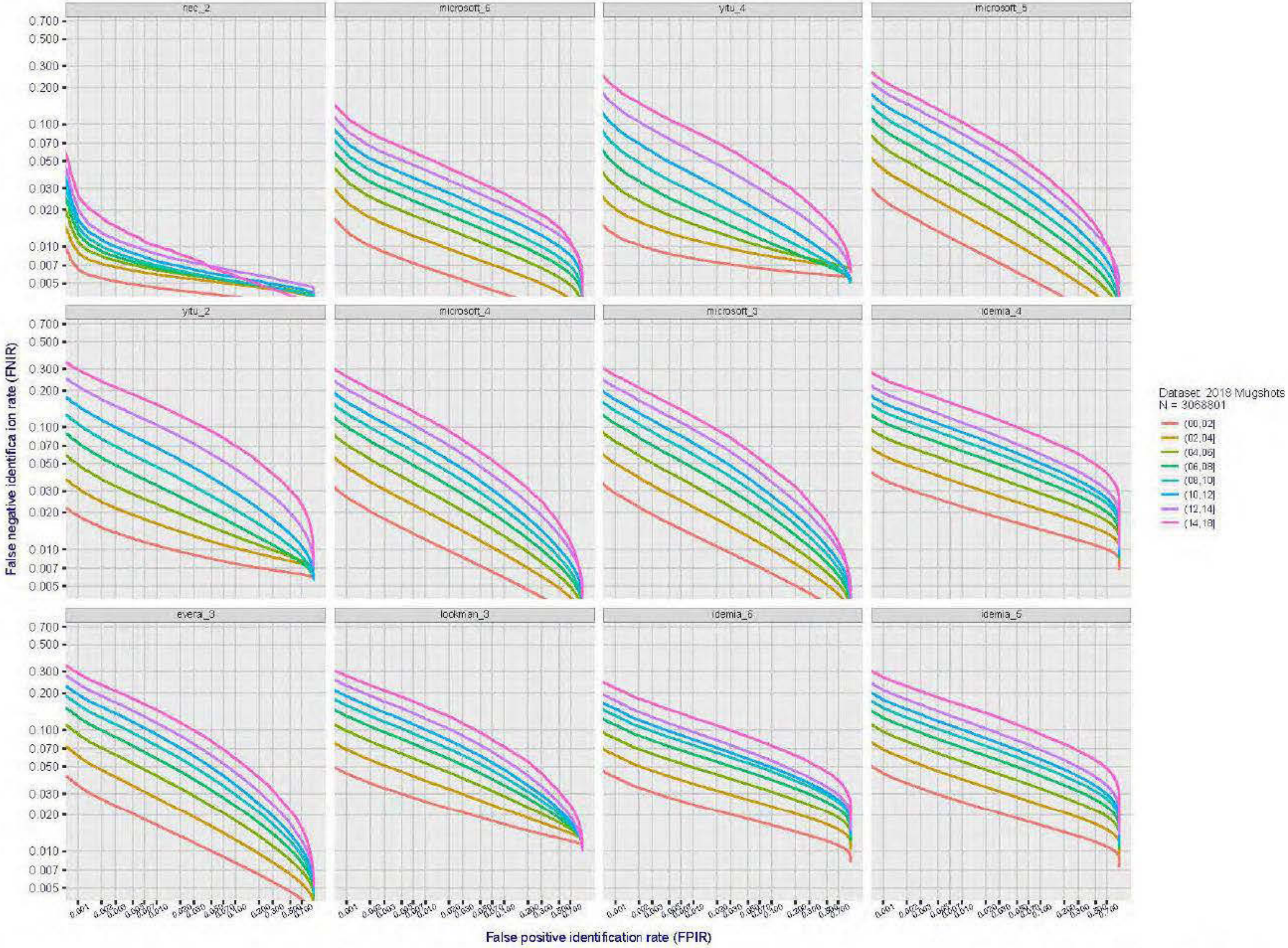


Figure 72: **[FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. FPIR by time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 5 and binned by number of years between search and initial enrollment. FPIR is computed from the same FRVT 2018 non-mates noted in row 3 of Table 5 with N = 3 000 000.*

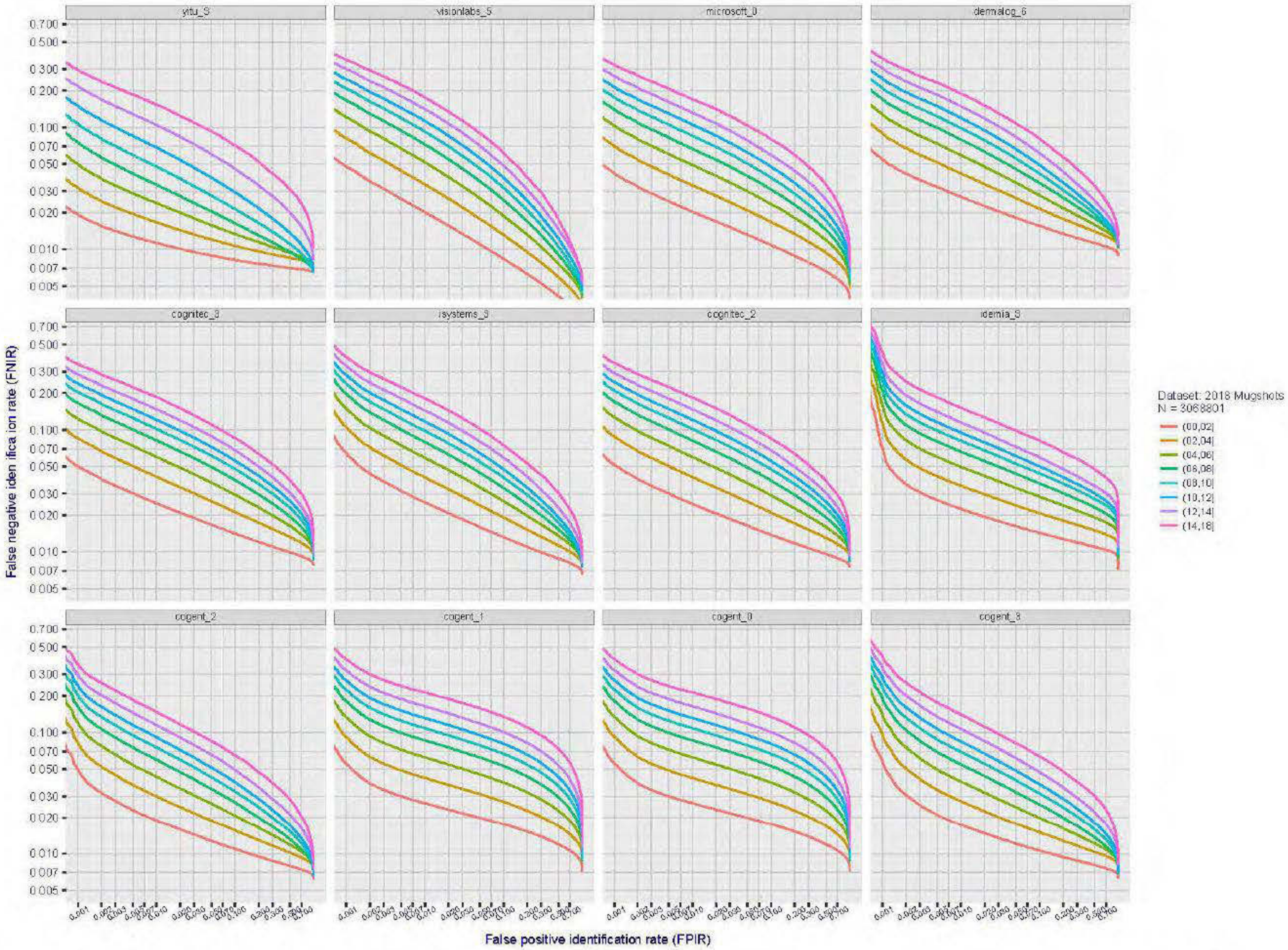


Figure 73: **[FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. FPIR by time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 5 and binned by number of years between search and initial enrollment. FPIR is computed from the same FRVT 2018 non-mates noted in row 3 of Table 5 with N = 3 000 000.*

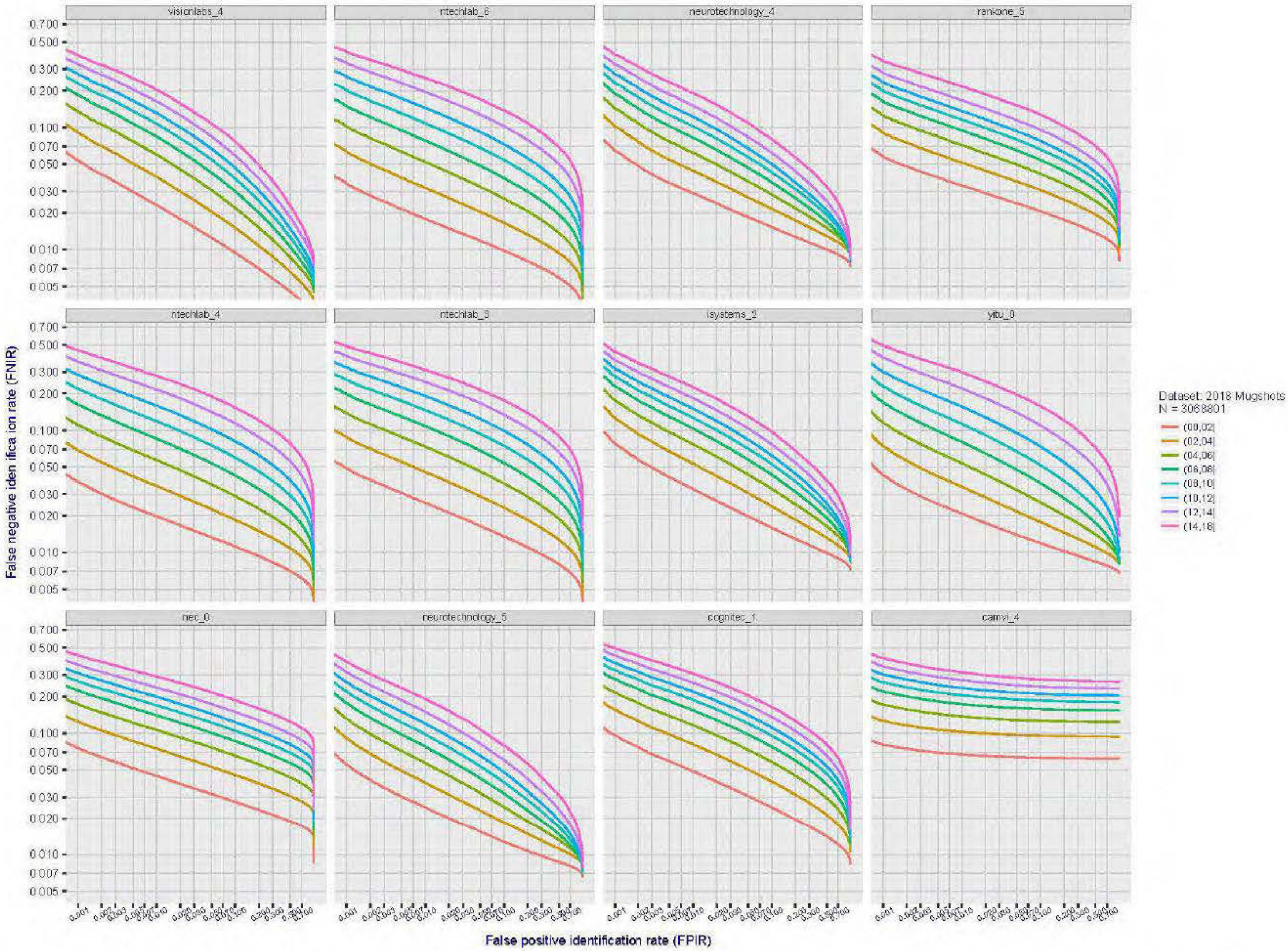


Figure 74: **[FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. FPIR by time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 5 and binned by number of years between search and initial enrollment. FPIR is computed from the same FRVT 2018 non-mates noted in row 3 of Table 5 with N = 3 000 000.*

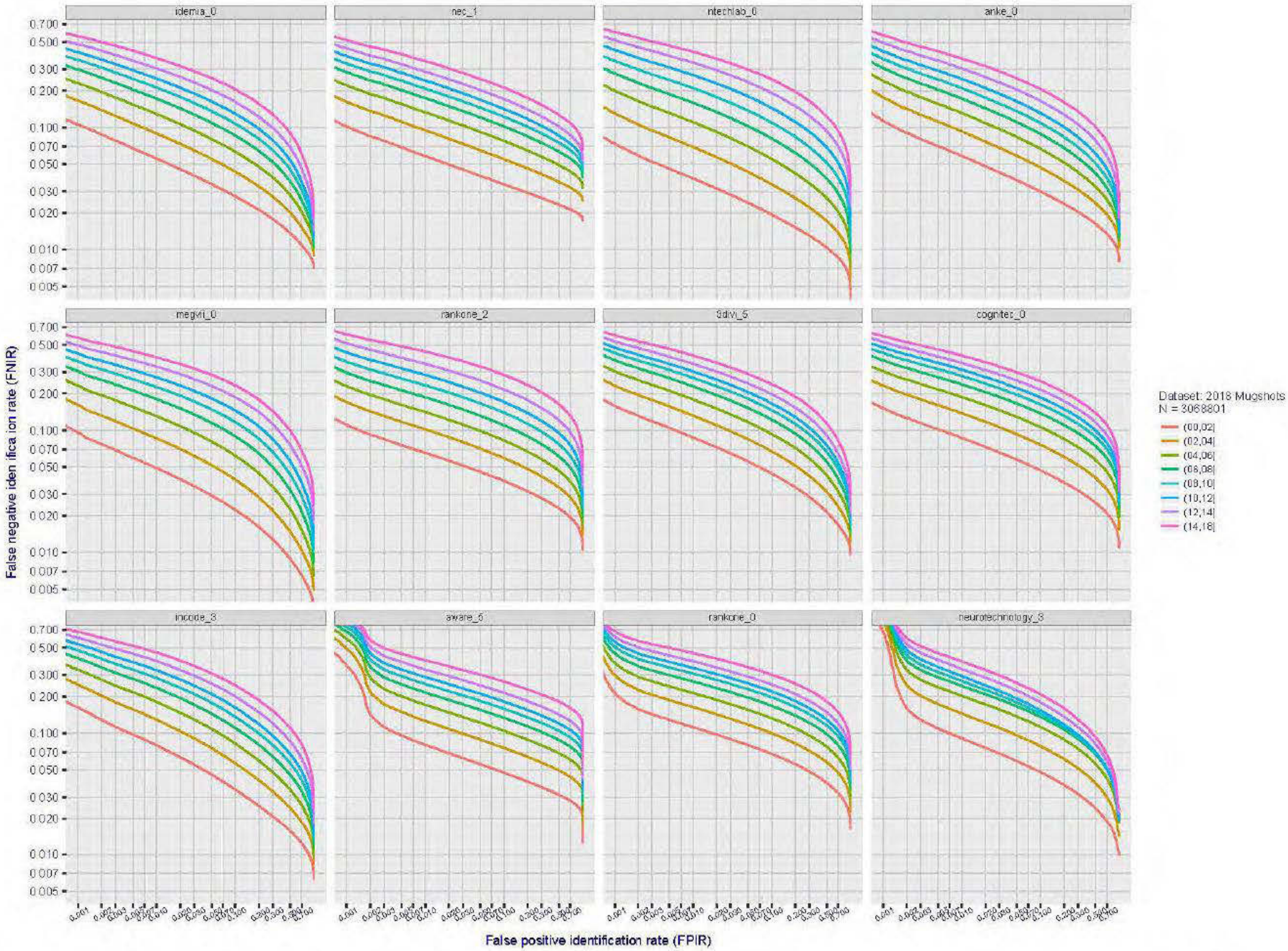


Figure 75: **[FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. FPIR by time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 5 and binned by number of years between search and initial enrollment. FPIR is computed from the same FRVT 2018 non-mates noted in row 3 of Table 5 with N = 3 000 000.*

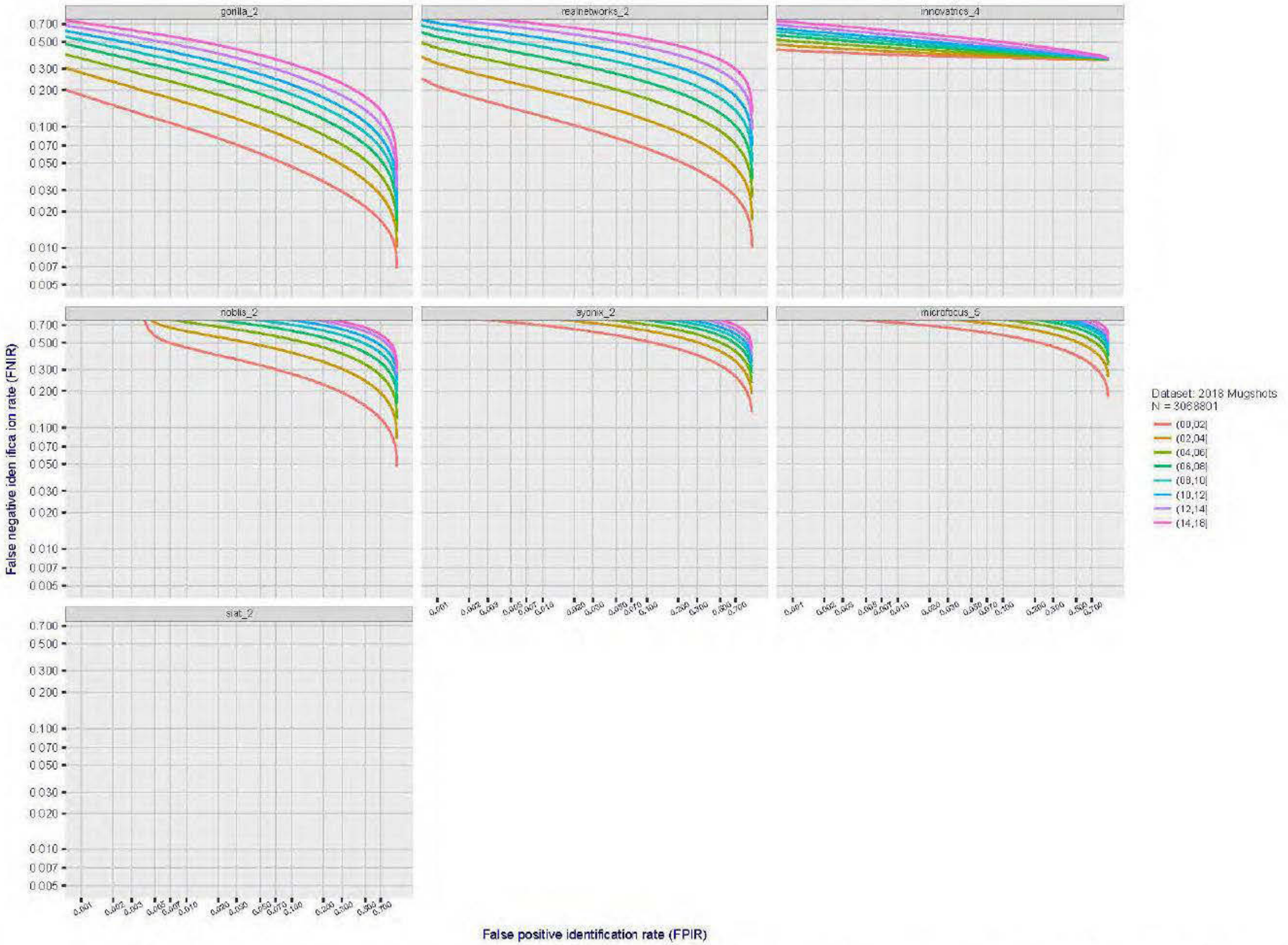


Figure 76: **[FRVT-2018 Mugshot Ageing Dataset] Identification miss rates vs. FPIR by time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Miss rates are computed over all searches noted in row 17 of Table 5 and binned by number of years between search and initial enrollment. FPIR is computed from the same FRVT 2018 non-mates noted in row 3 of Table 5 with N = 3 000 000.*

2019/09/11	FNIR(N, R, T) =	False neg. identification rate	N = Num. enrolled subjects	T = Threshold	T = 0 → Investigation
17:24:52	FPIR(N, T) =	False pos. identification rate	R = Num. candidates examined		T > 0 → Identification

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

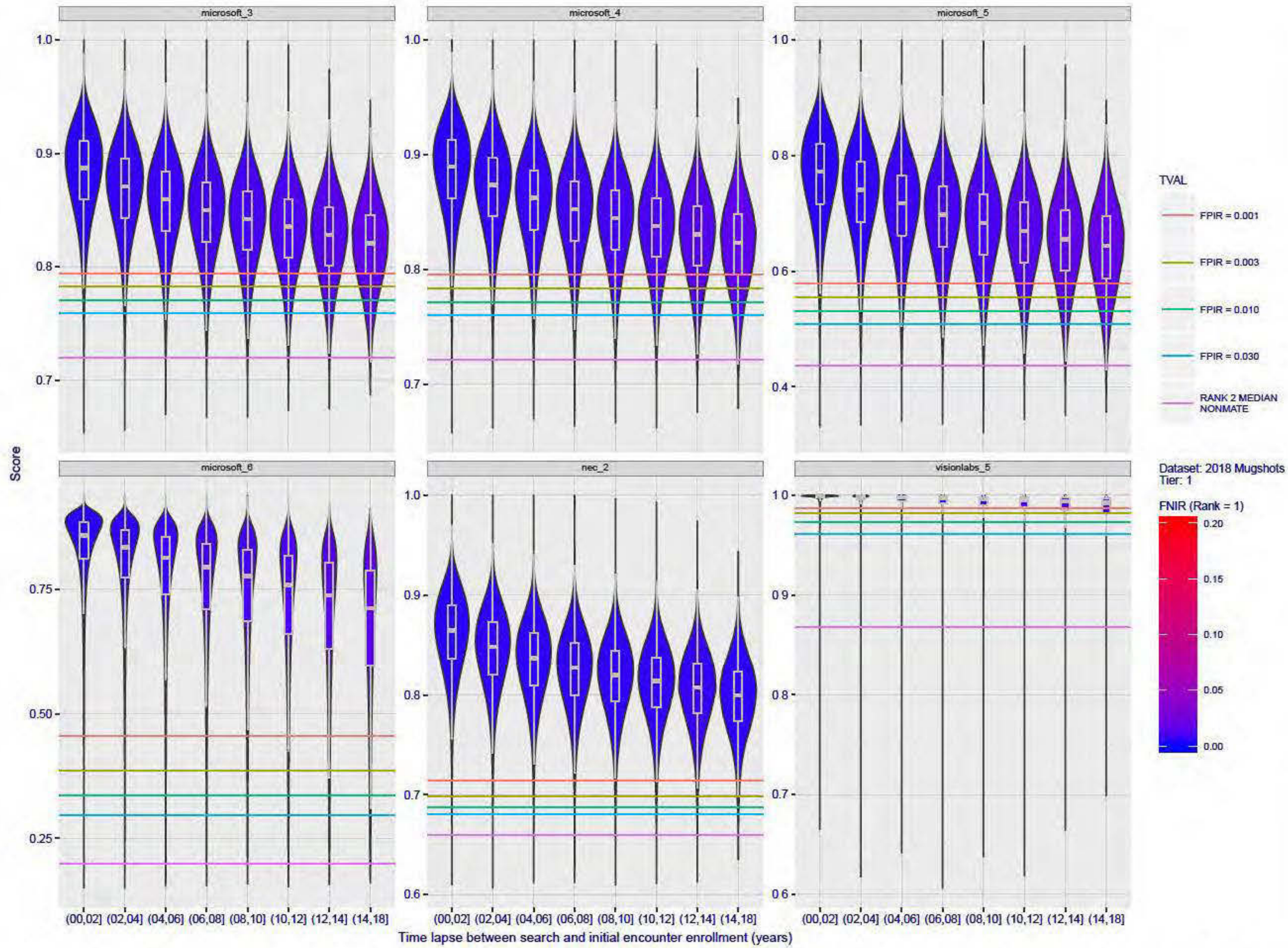


Figure 77: **[FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 5 binned by number of years between search and initial enrollment.*

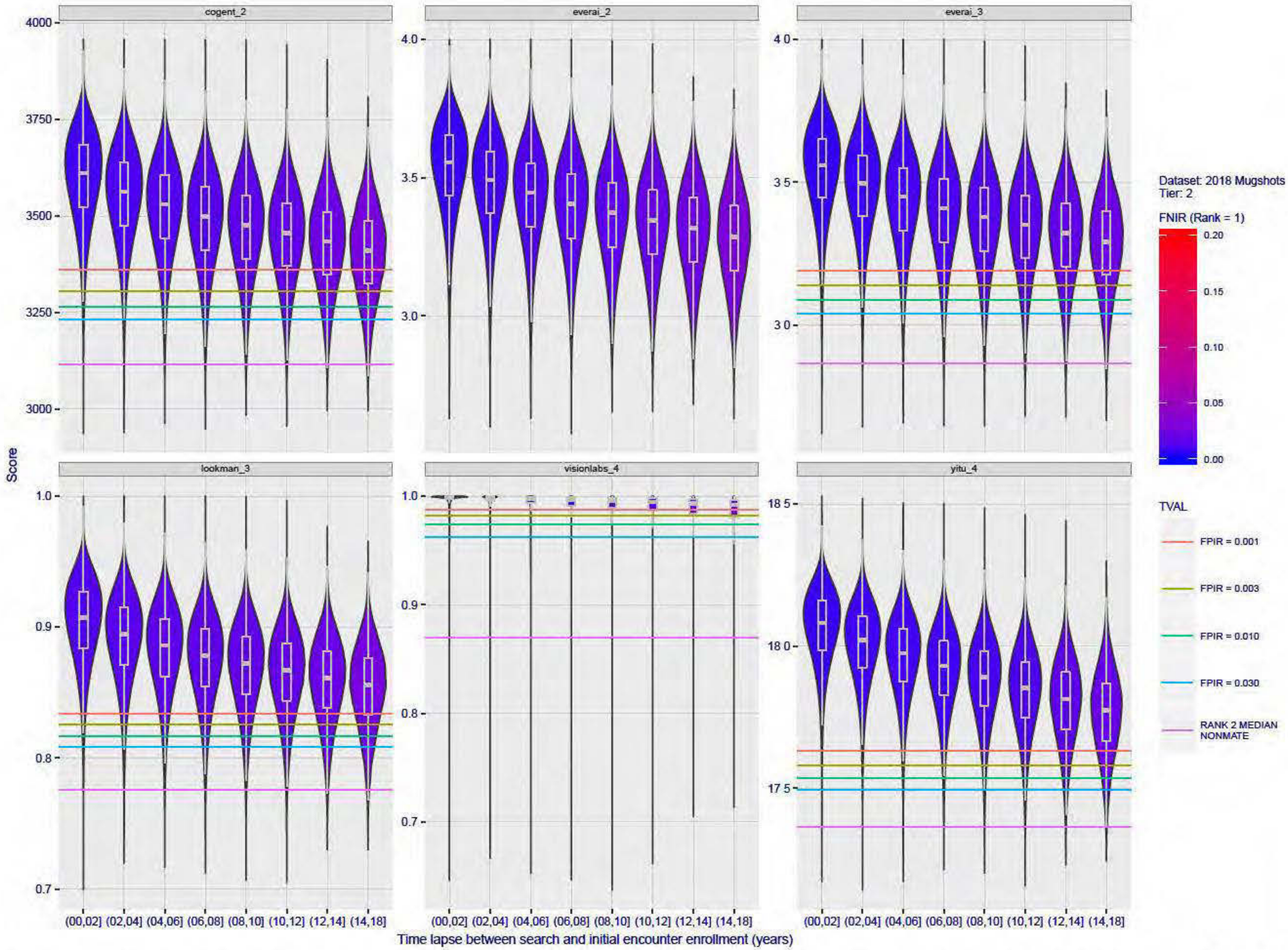


Figure 78: **[FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 5 binned by number of years between search and initial enrollment.*

FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

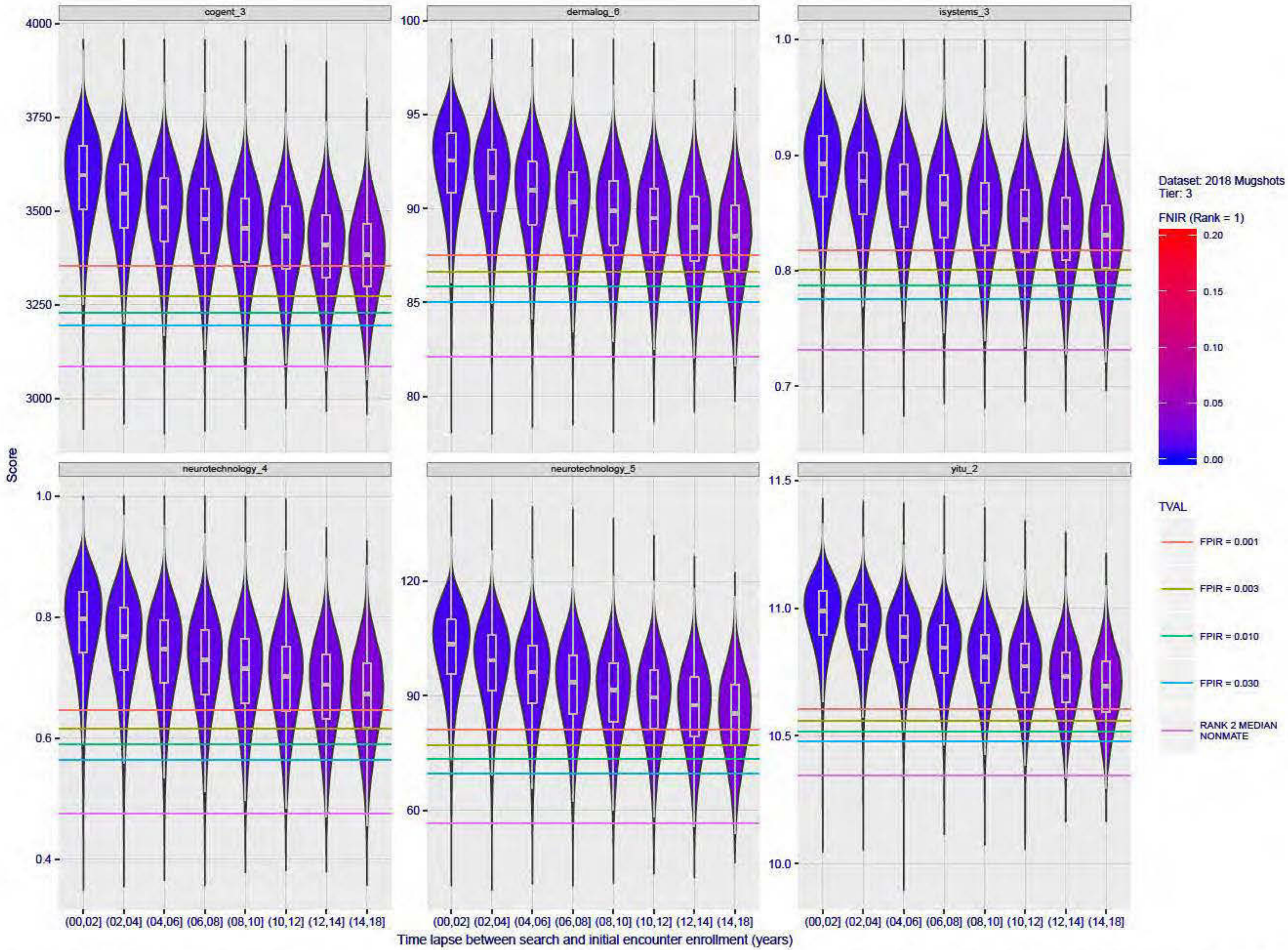


Figure 79: **[FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 5 binned by number of years between search and initial enrollment.*

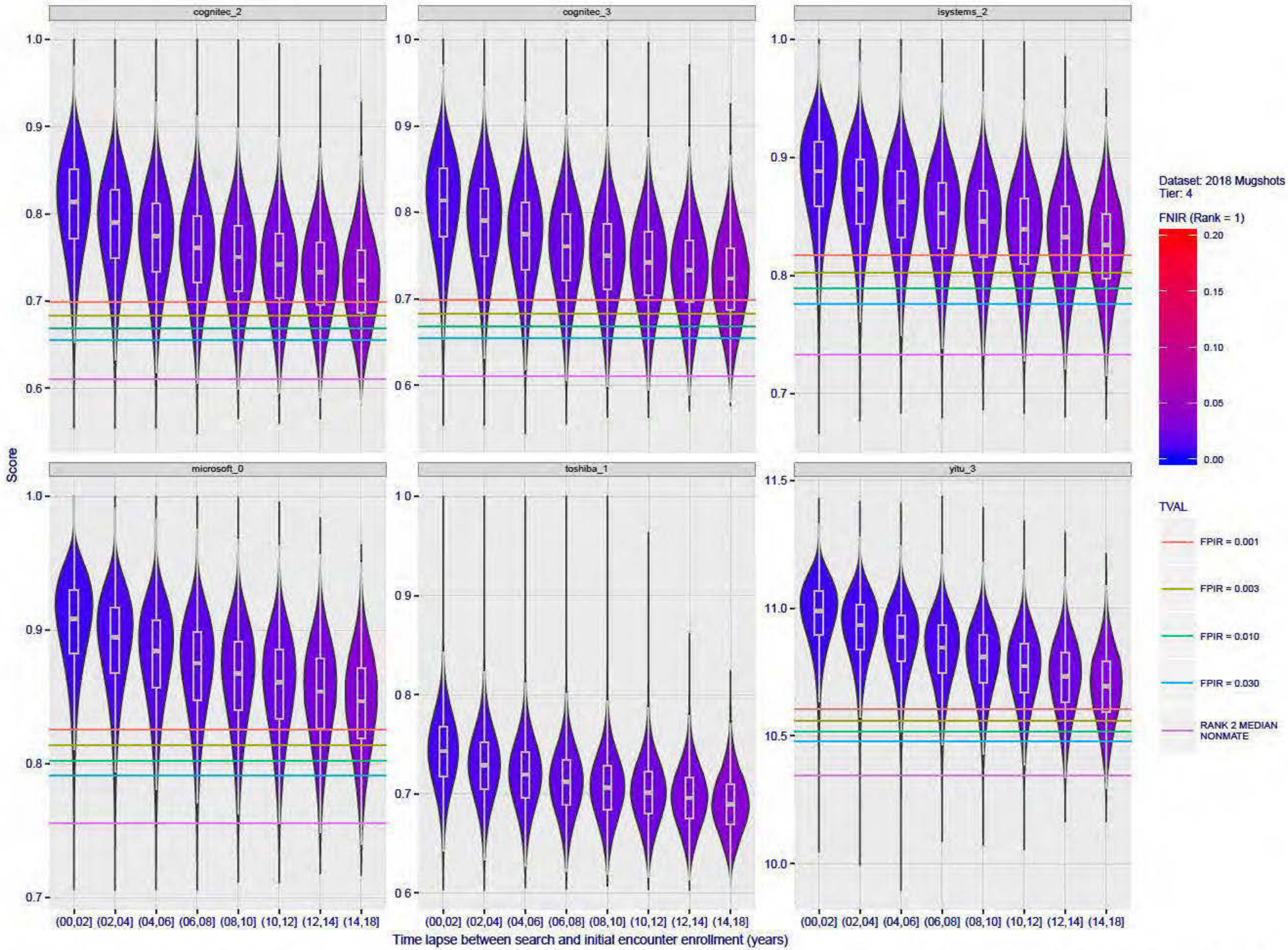


Figure 80: **[FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 5 binned by number of years between search and initial enrollment.*

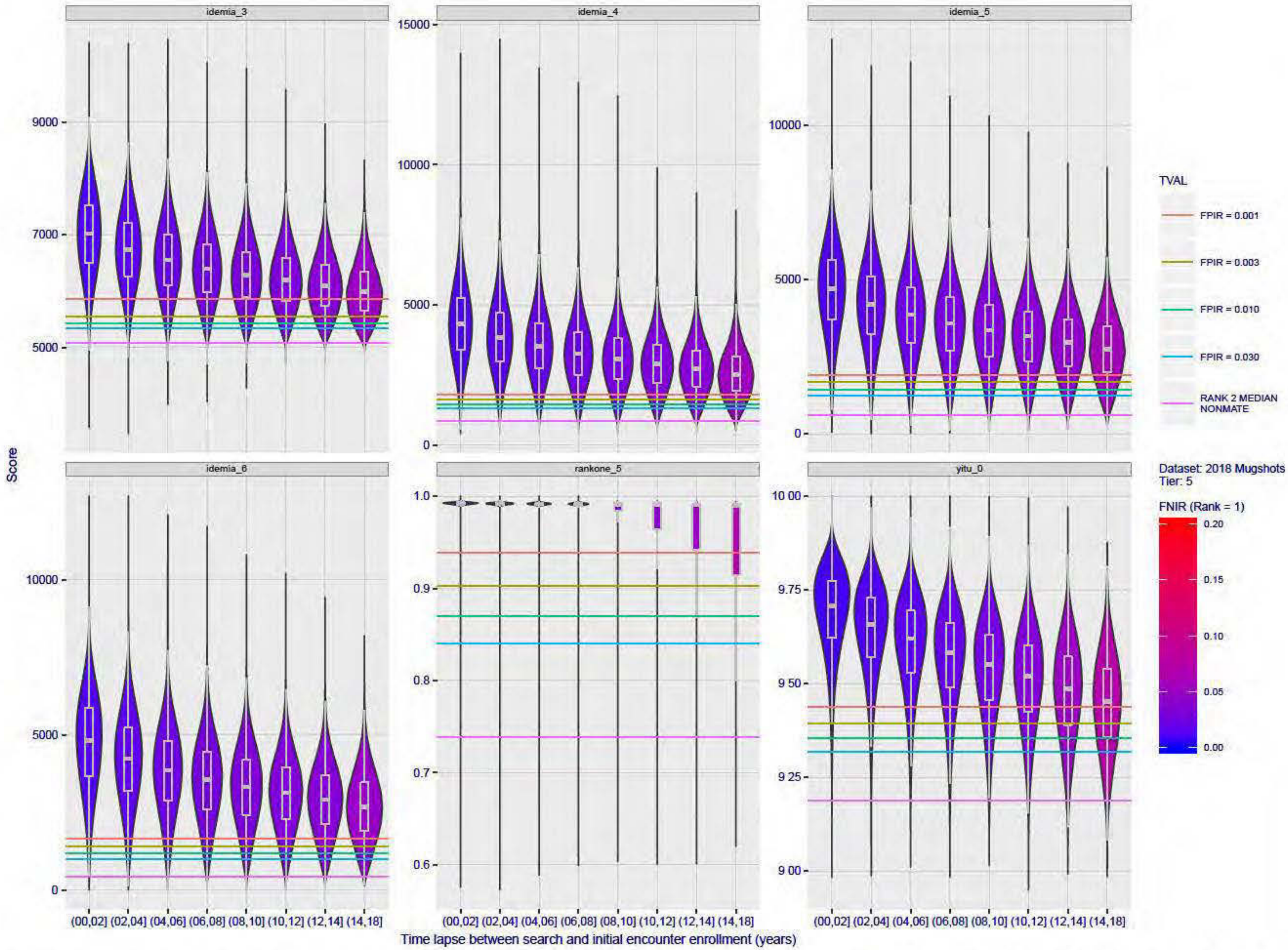


Figure 81: **[FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 5 binned by number of years between search and initial enrollment.*

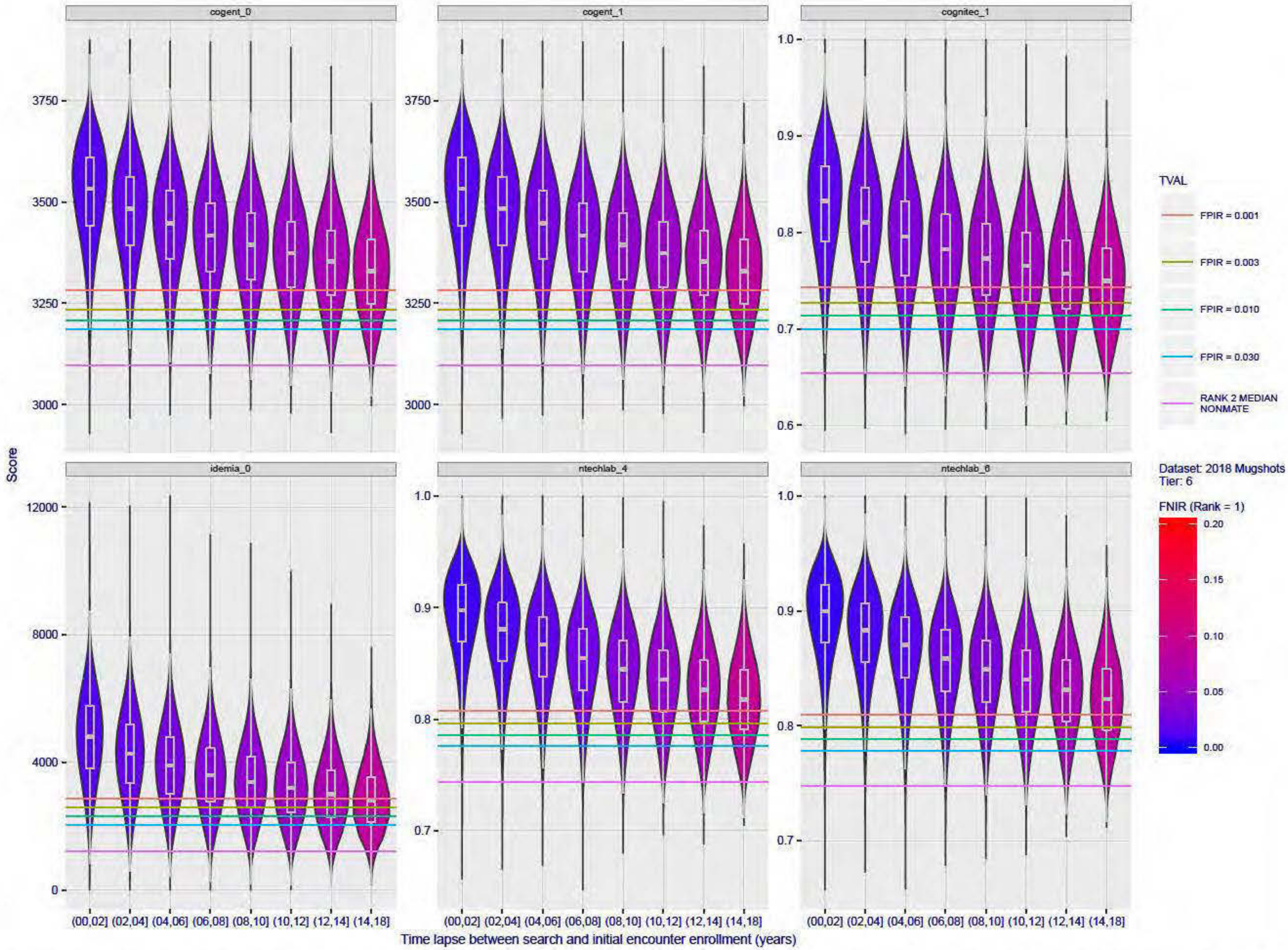


Figure 82: **[FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 5 binned by number of years between search and initial enrollment.*

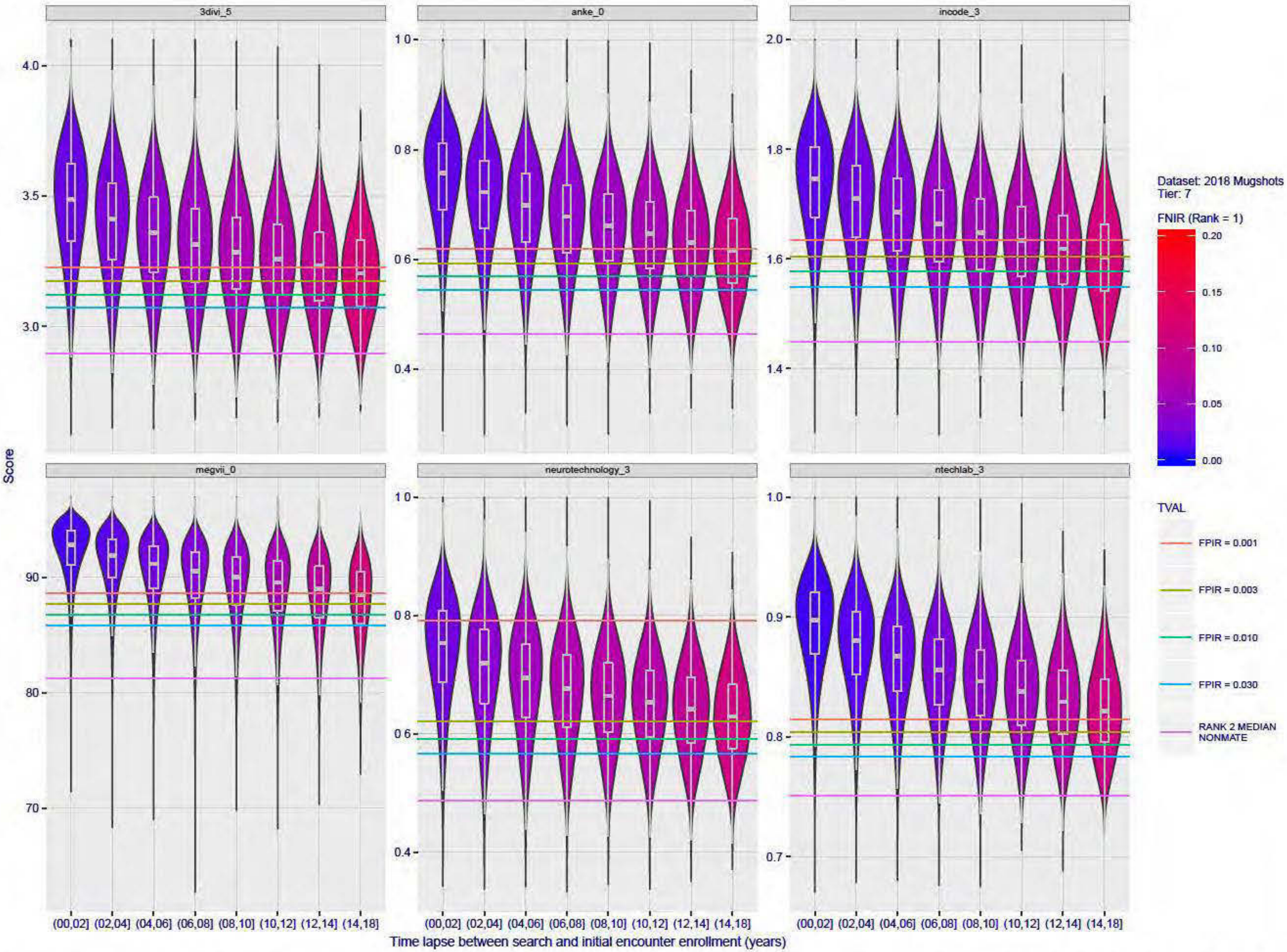


Figure 83: **[FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 5 binned by number of years between search and initial enrollment.*

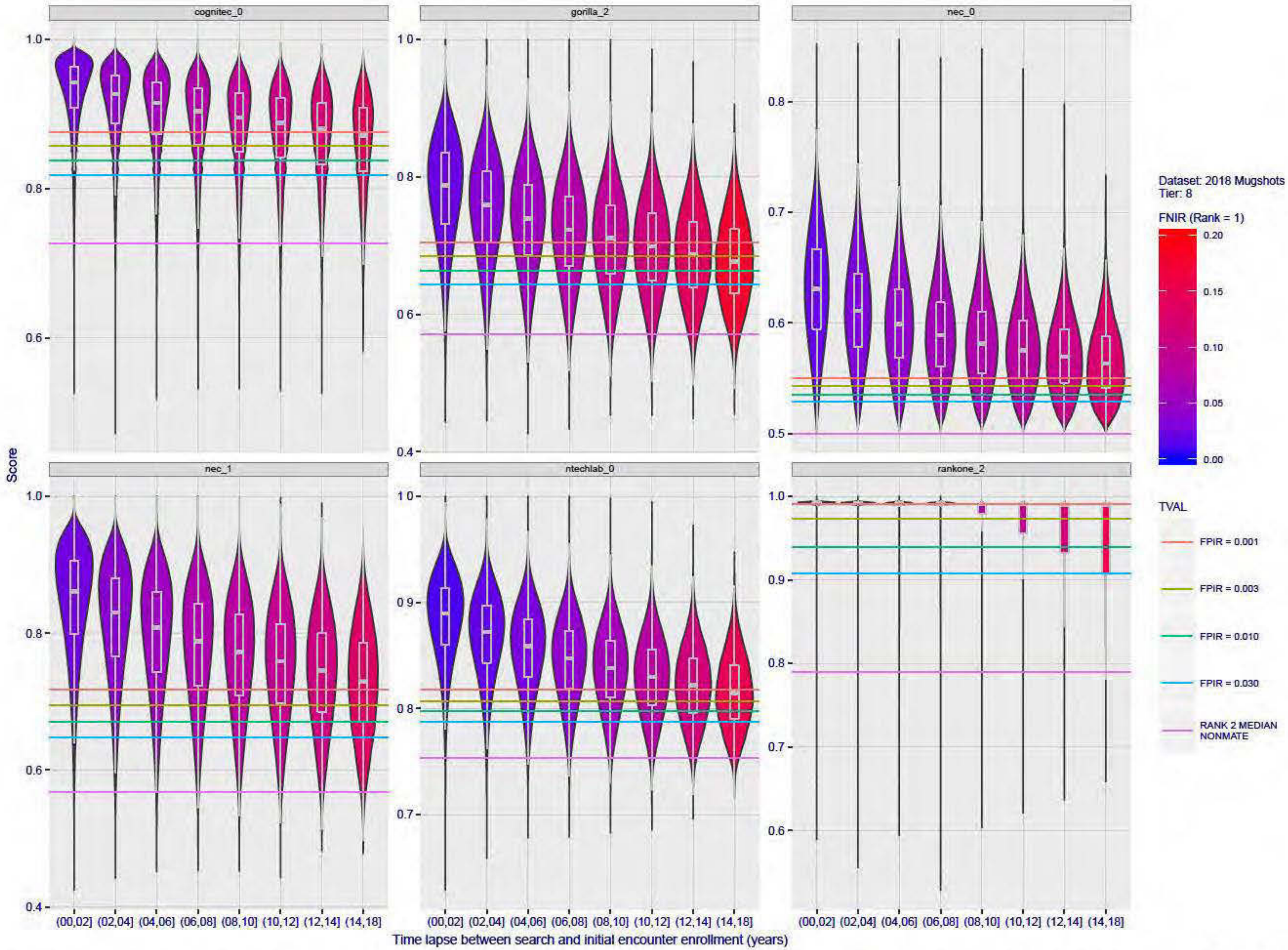


Figure 84: **[FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 5 binned by number of years between search and initial enrollment.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

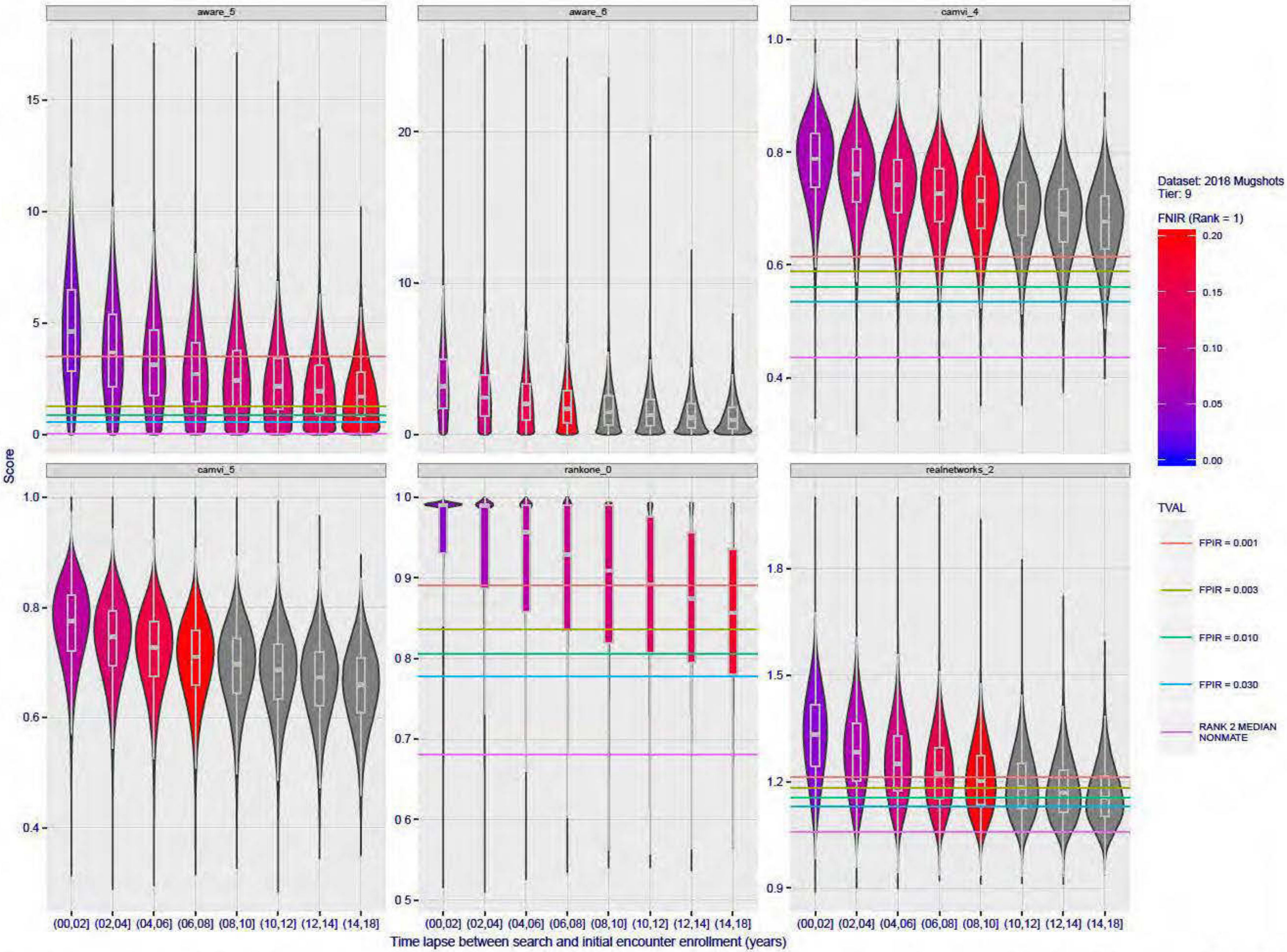


Figure 85: **[FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 5 binned by number of years between search and initial enrollment.*

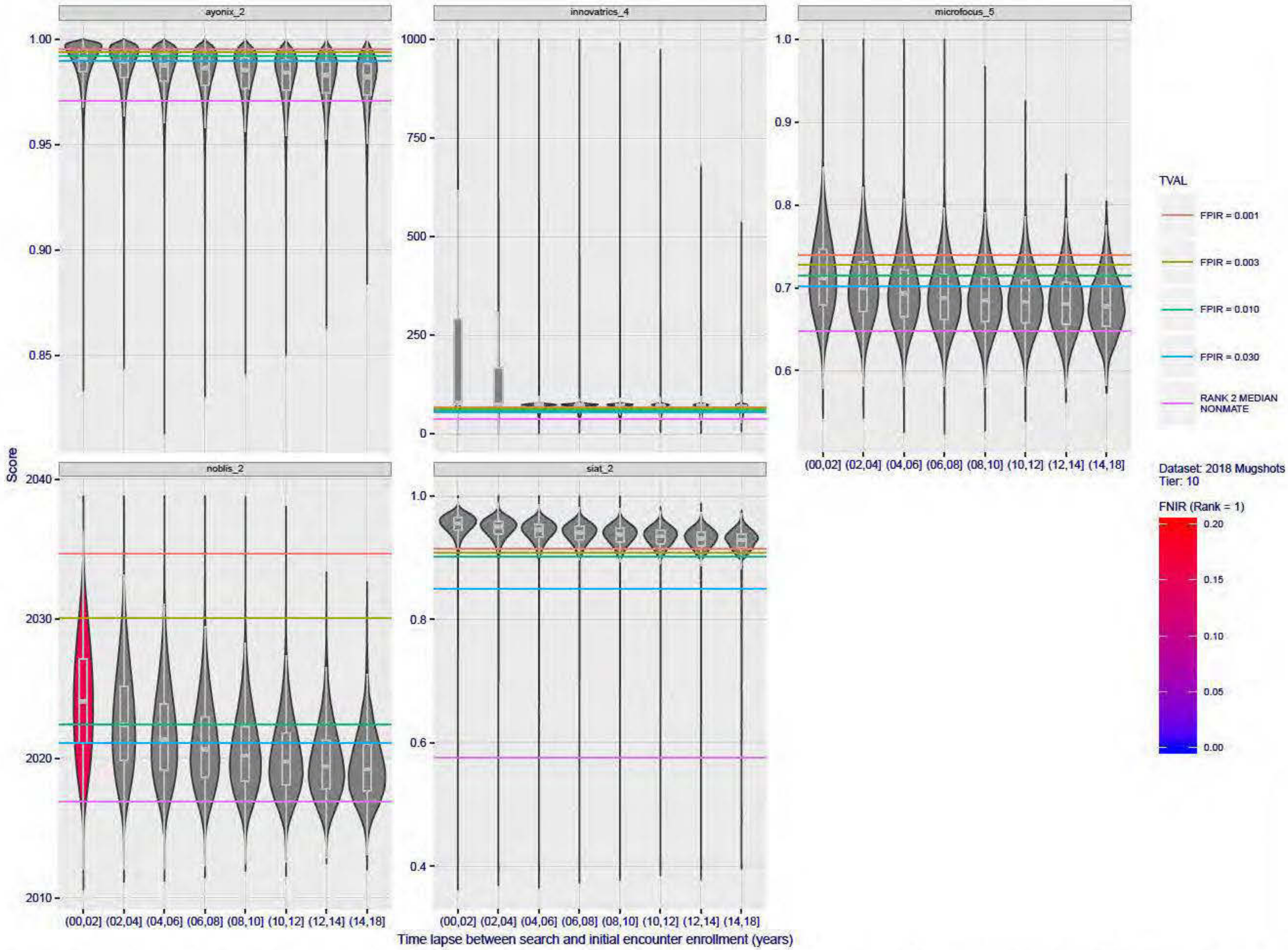


Figure 86: **[FRVT-2018 Mugshot Ageing Dataset] Native mate scores vs. time-elapsed.** *The oldest image of each individual is enrolled. Thereafter, all more recent images are searched. Mated score distributions are computed over all searches noted in row 17 of Table 5 binned by number of years between search and initial enrollment.*

Appendix C Effect of enrolling multiple images

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

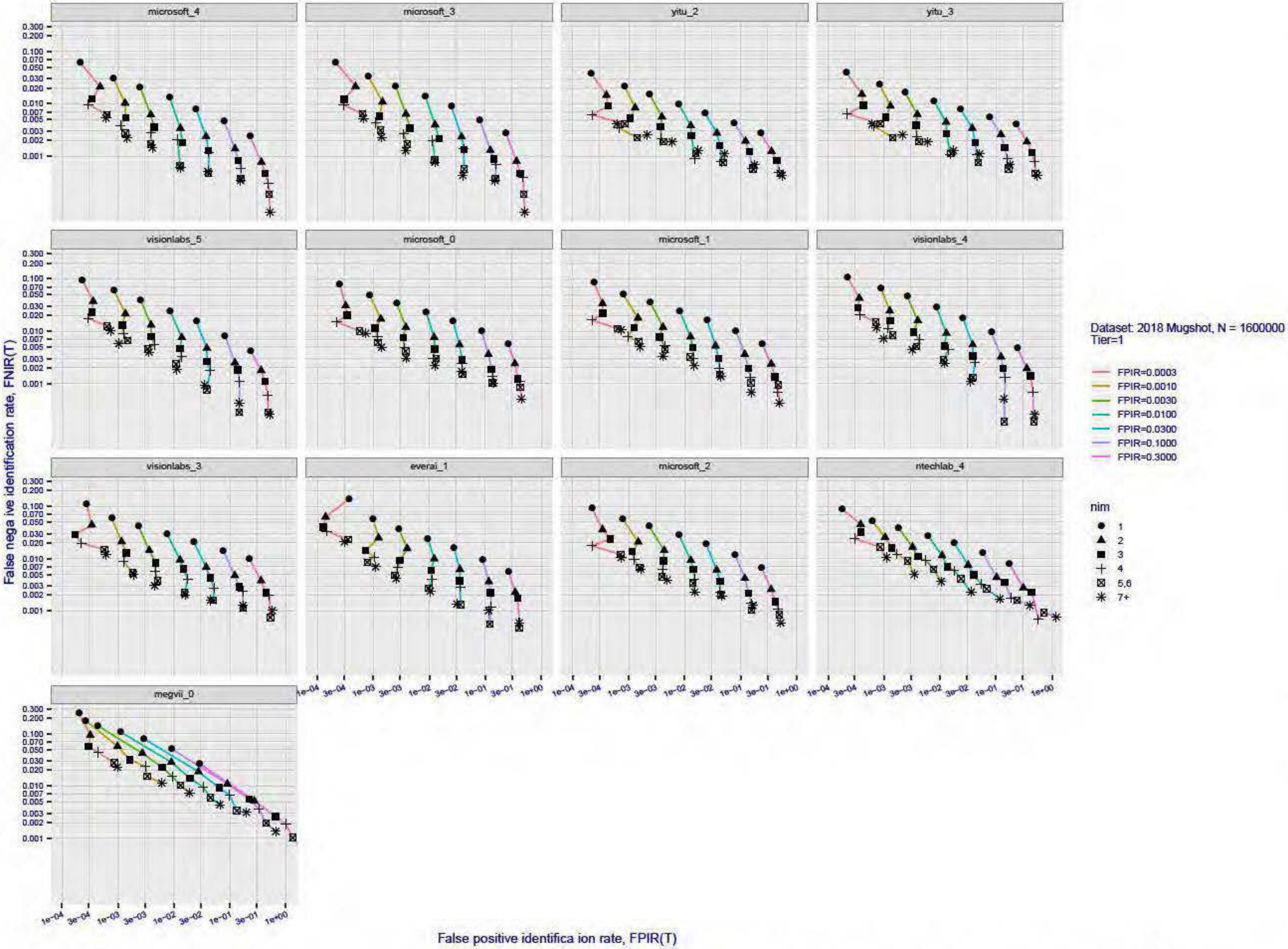


Figure 87: **[FRVT-2018 Mugshot Dataset] Effect of enrolling multiple images for each identity.** *The plot shows an identification miss rates vs. false positive rates, at seven operating thresholds. The enrolled population size is fixed. The images are enrolled with lifetime-consolidation - see section 2.3.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

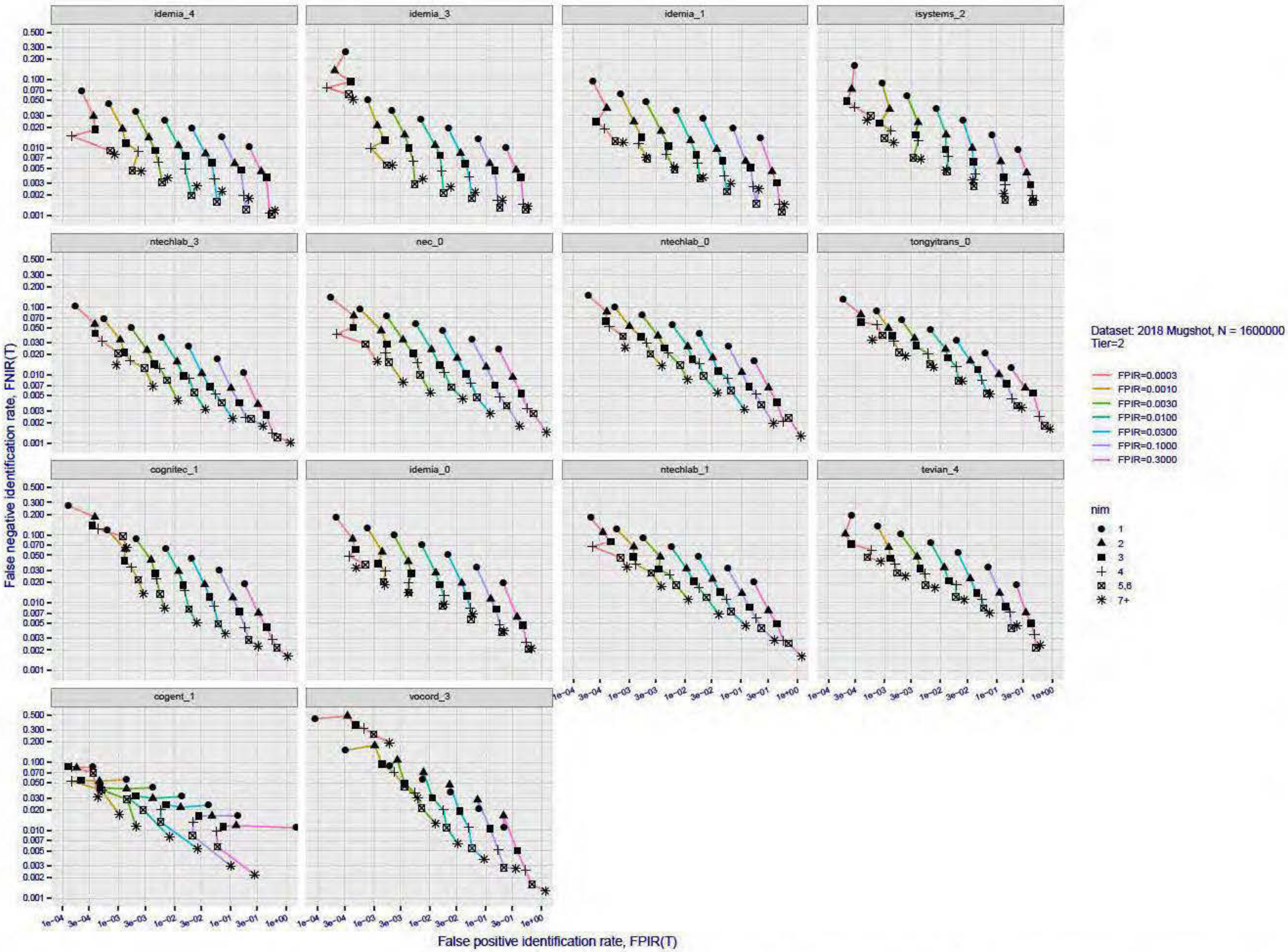


Figure 88: **[FRVT-2018 Mugshot Dataset] Effect of enrolling multiple images for each identity.** *The plot shows an identification miss rates vs. false positive rates, at seven operating thresholds. The enrolled population size is fixed. The images are enrolled with lifetime-consolidation - see section 2.3.*

2019/09/11 17:24:52

FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

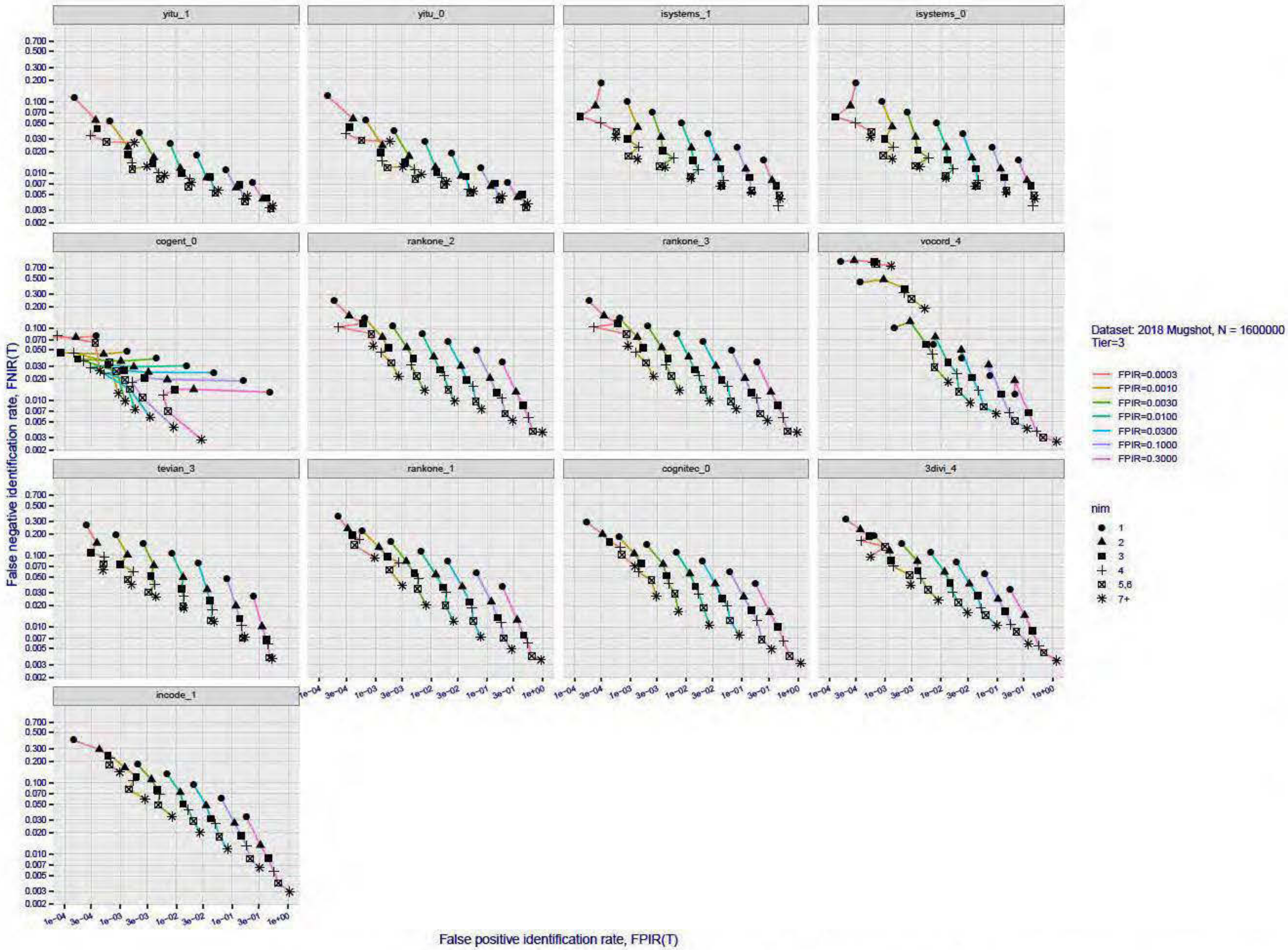


Figure 89: **[FRVT-2018 Mugshot Dataset] Effect of enrolling multiple images for each identity.** *The plot shows an identification miss rates vs. false positive rates, at seven operating thresholds. The enrolled population size is fixed. The images are enrolled with lifetime-consolidation - see section 2.3.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

2019/09/11 17:24:52

FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

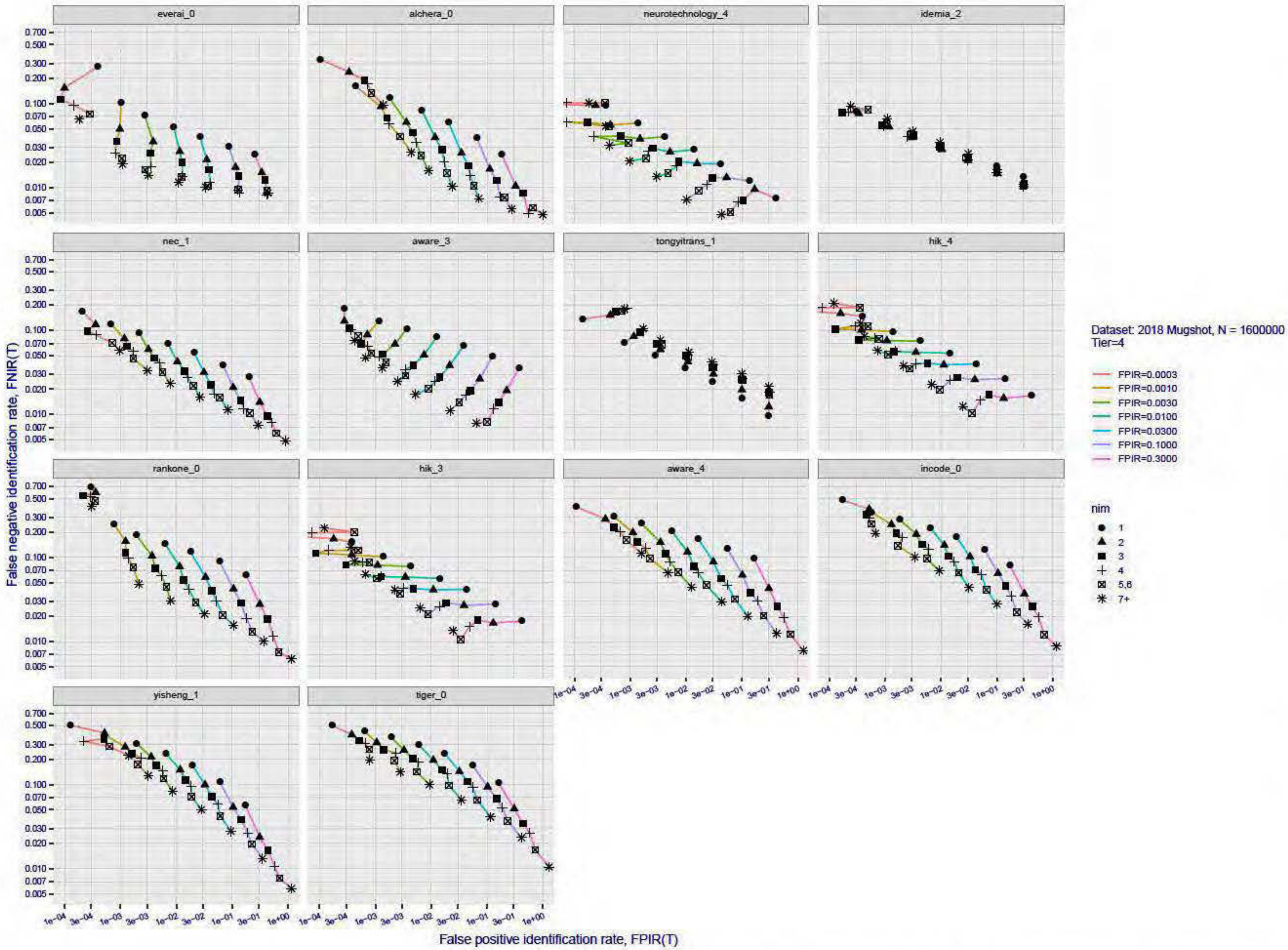


Figure 90: **[FRVT-2018 Mugshot Dataset] Effect of enrolling multiple images for each identity.** *The plot shows an identification miss rates vs. false positive rates, at seven operating thresholds. The enrolled population size is fixed. The images are enrolled with lifetime-consolidation - see section 2.3.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

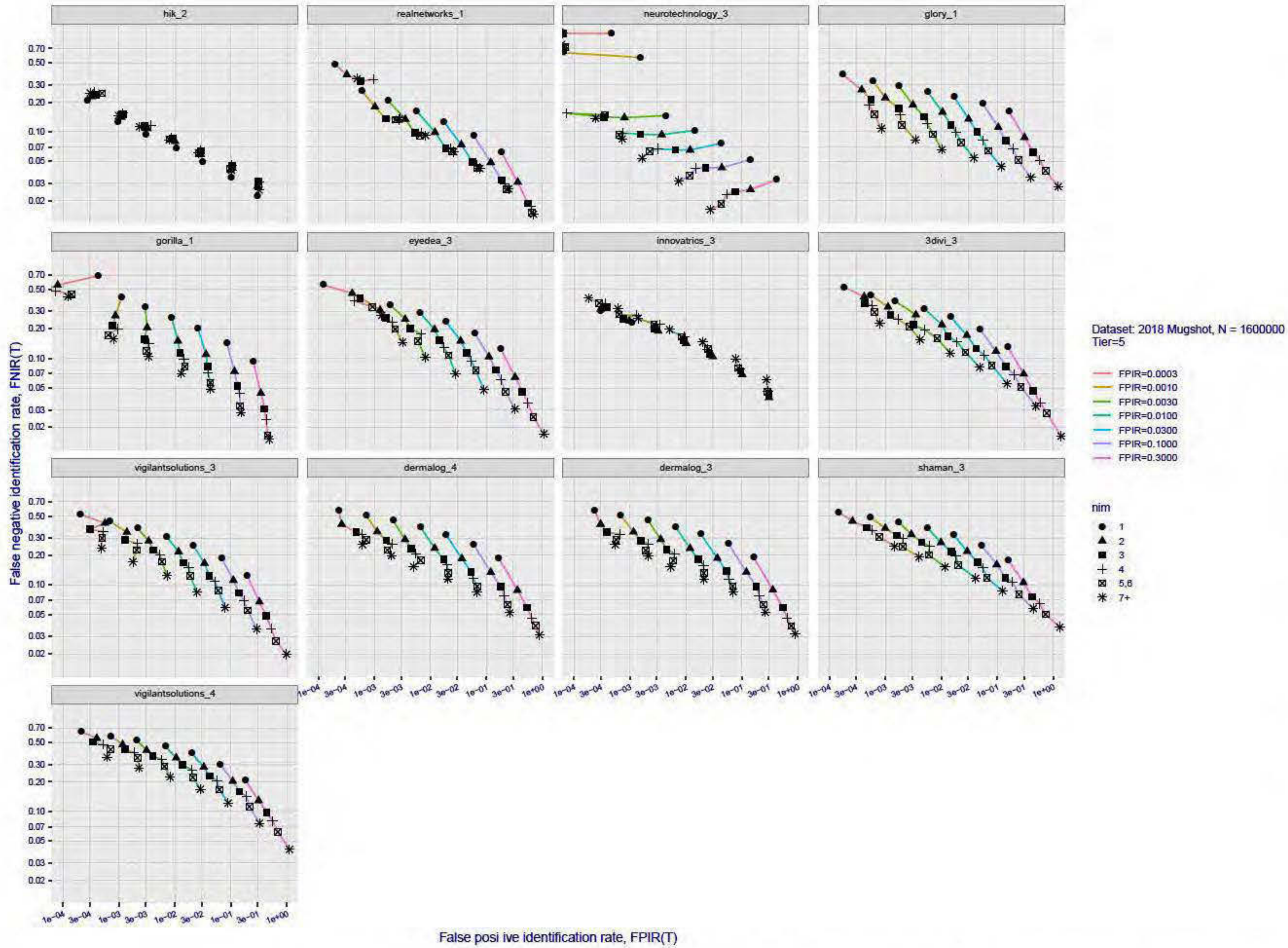


Figure 91: **[FRVT-2018 Mugshot Dataset] Effect of enrolling multiple images for each identity.** *The plot shows an identification miss rates vs. false positive rates, at seven operating thresholds. The enrolled population size is fixed. The images are enrolled with lifetime-consolidation - see section 2.3.*

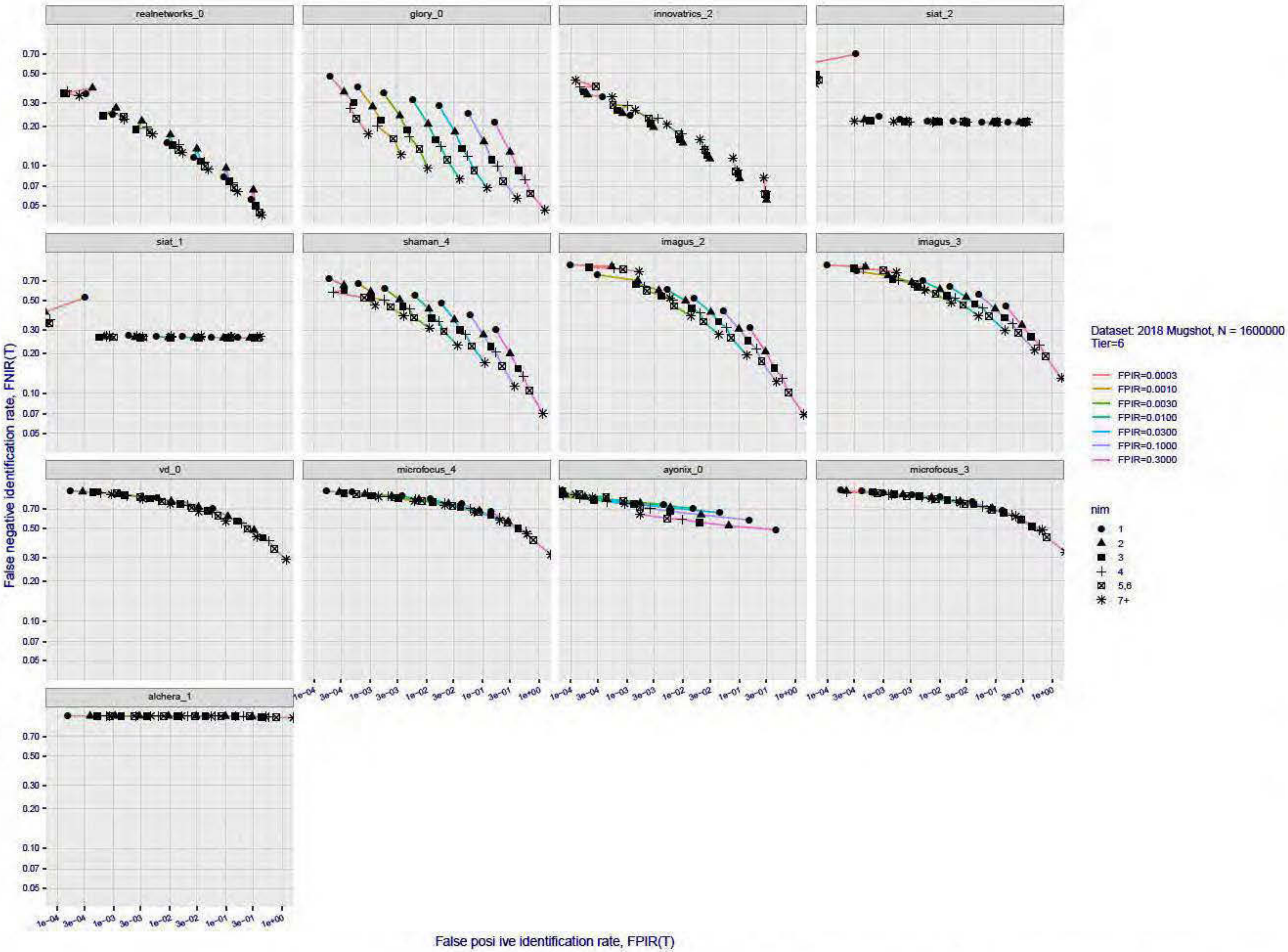


Figure 92: **[FRVT-2018 Mugshot Dataset] Effect of enrolling multiple images for each identity.** *The plot shows an identification miss rates vs. false positive rates, at seven operating thresholds. The enrolled population size is fixed. The images are enrolled with lifetime-consolidation - see section 2.3.*

Appendix D Accuracy with poor quality webcam images

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

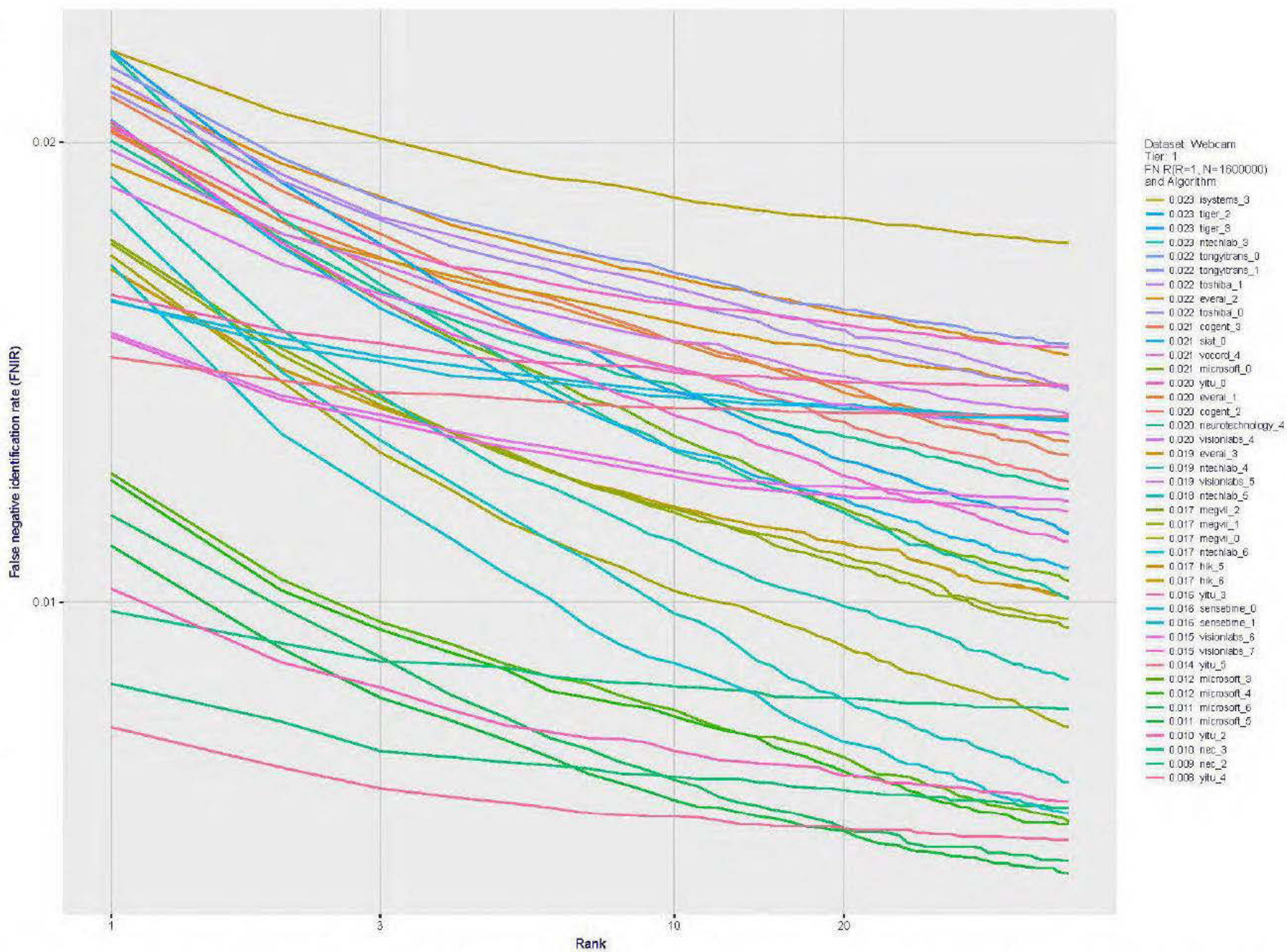


Figure 93: **[Webcam Dataset] Identification miss rates vs. rank.** *The results apply to cross-domain recognition in which webcams are searched against enrolled mugshots. The FNIR values are higher than those for mugshot-mugshot identification due to low image resolution, lighting and less constrained subject pose in webcam images - see Figure 4.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

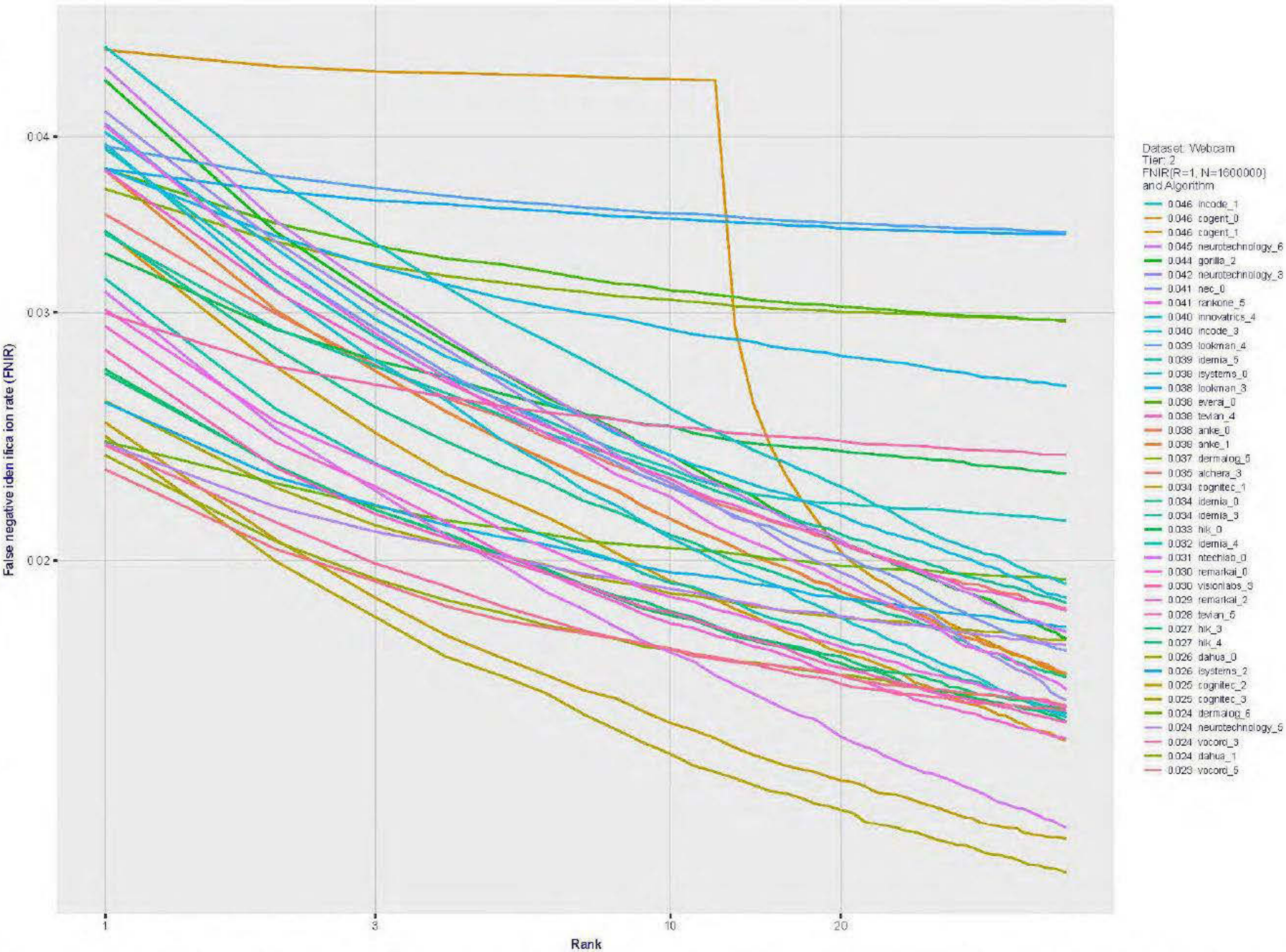


Figure 94: **[Webcam Dataset] Identification miss rates vs. rank.** *The results apply to cross-domain recognition in which webcams are searched against enrolled mugshots. The FNIR values are higher than those for mugshot-mugshot identification due to low image resolution, lighting and less constrained subject pose in webcam images - see Figure 4.*

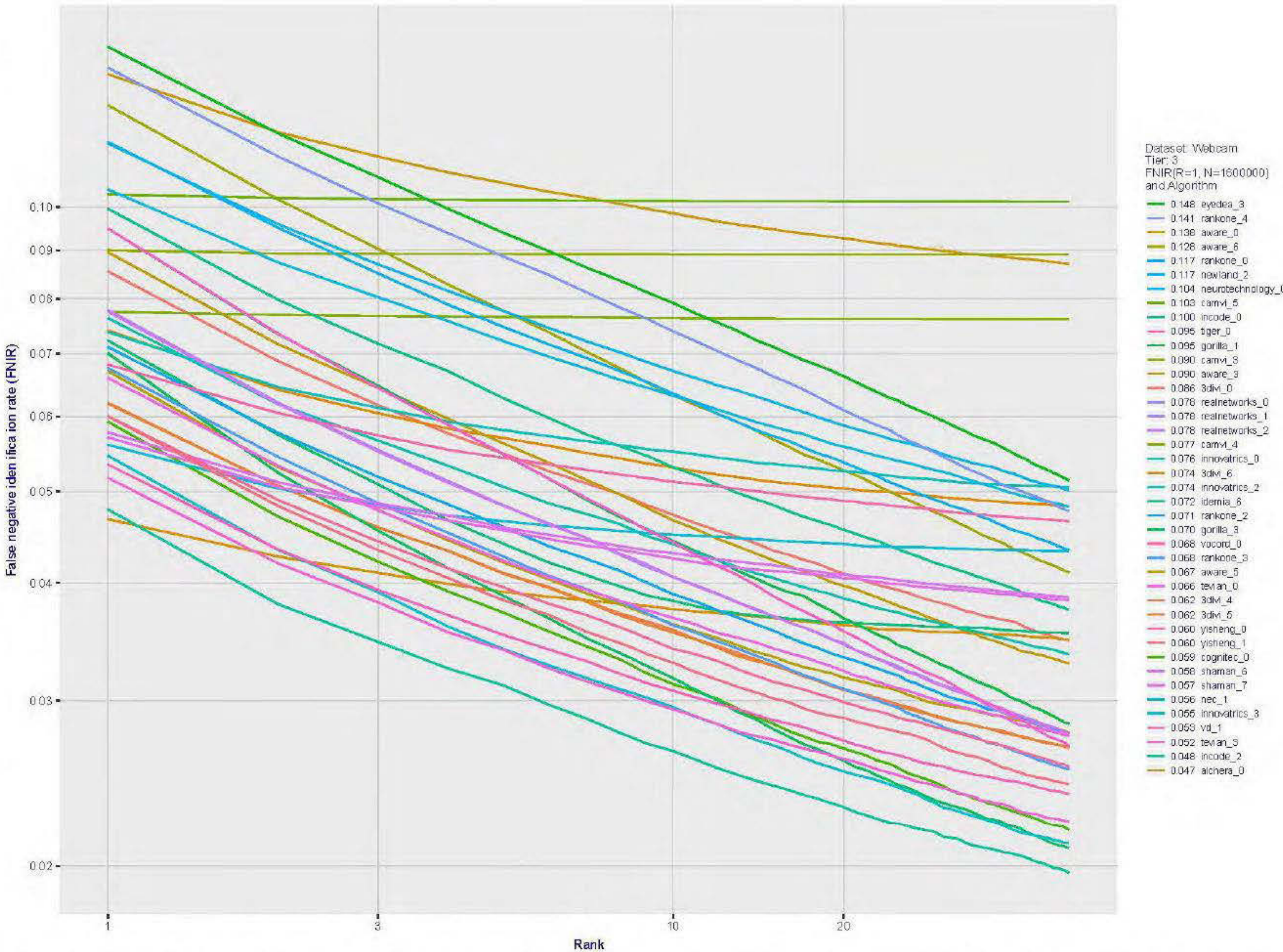


Figure 95: **[Webcam Dataset] Identification miss rates vs. rank.** *The results apply to cross-domain recognition in which webcams are searched against enrolled mugshots. The FNIR values are higher than those for mugshot-mugshot identification due to low image resolution, lighting and less constrained subject pose in webcam images - see Figure 4.*

FNIR(N, R, T) = False neg. identification rate
FPIR(N, T) = False pos. identification rate
N = Num. enrolled subjects
R = Num. candidates examined
T = Threshold
T = 0 → Investigation
T > 0 → Identification

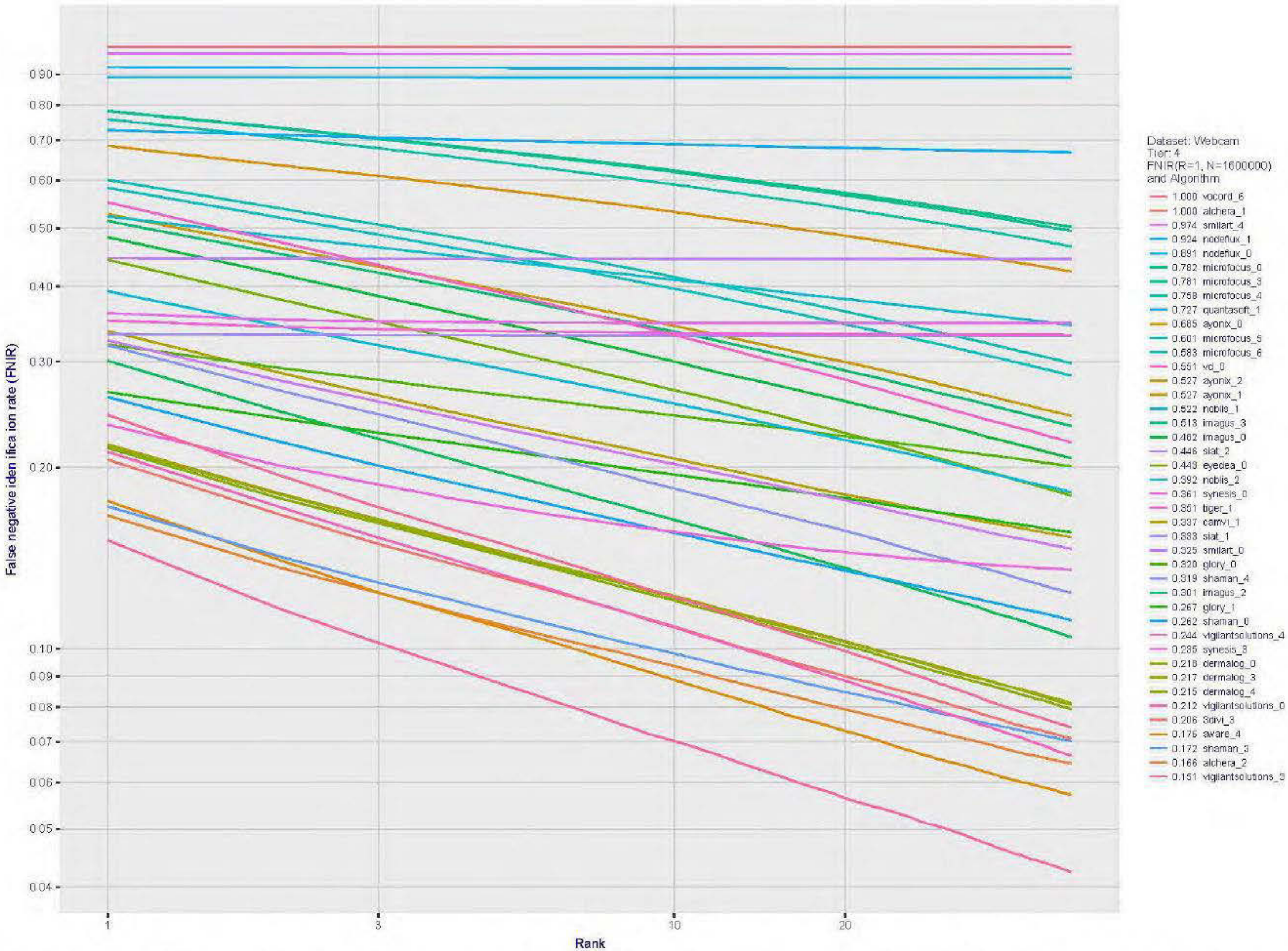


Figure 96: **[Webcam Dataset] Identification miss rates vs. rank.** *The results apply to cross-domain recognition in which webcams are searched against enrolled mugshots. The FNIR values are higher than those for mugshot-mugshot identification due to low image resolution, lighting and less constrained subject pose in webcam images - see Figure 4.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

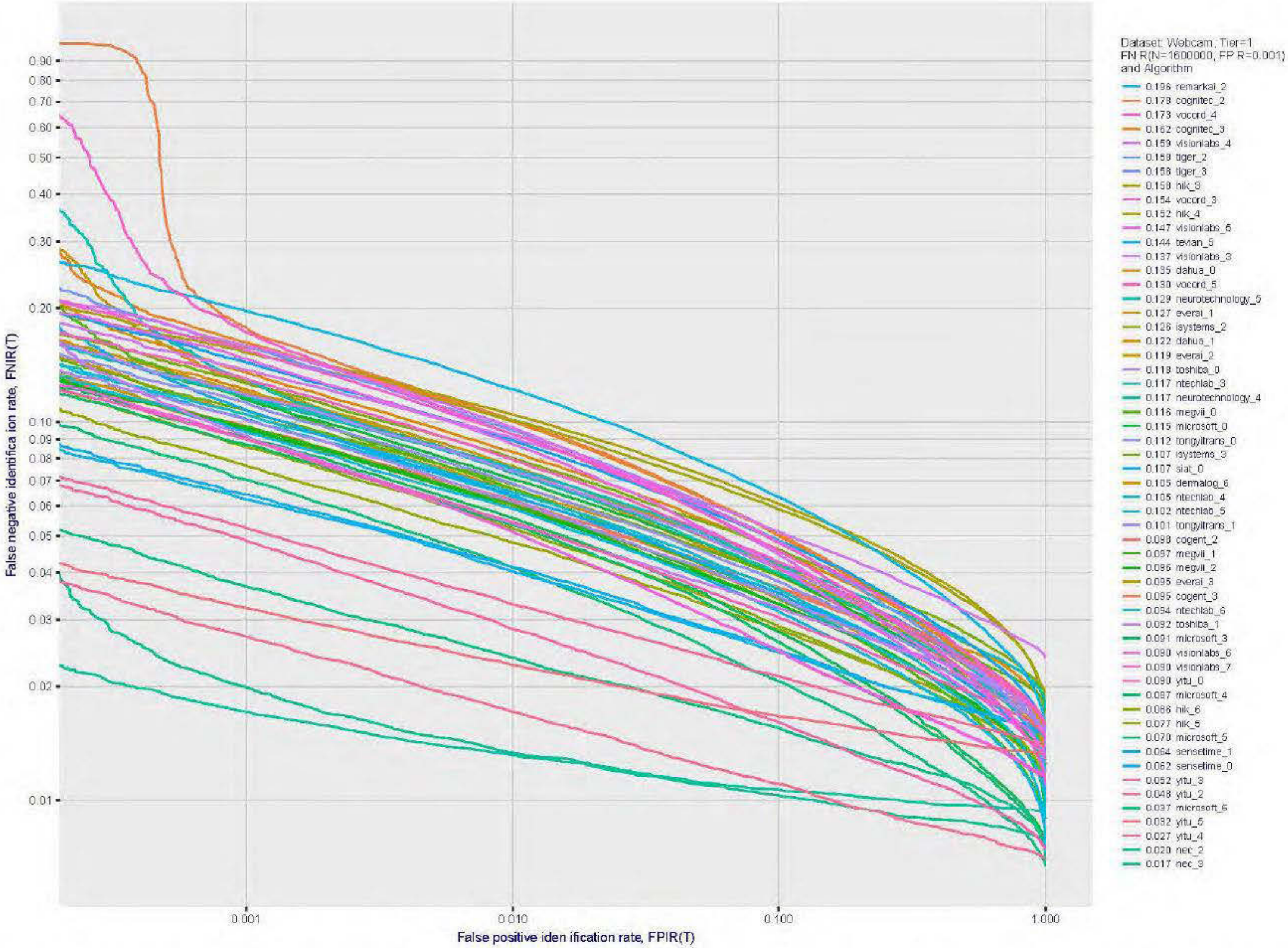


Figure 97: **[Webcam Dataset] Identification miss rates vs. false positive rates.** *The results apply to cross-domain recognition in which webcams are searched against enrolled mugshots. The FNIR values are higher than those for mugshot-mugshot identification due to low image resolution, lighting and less constrained subject pose in webcam images - see Figure 4.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

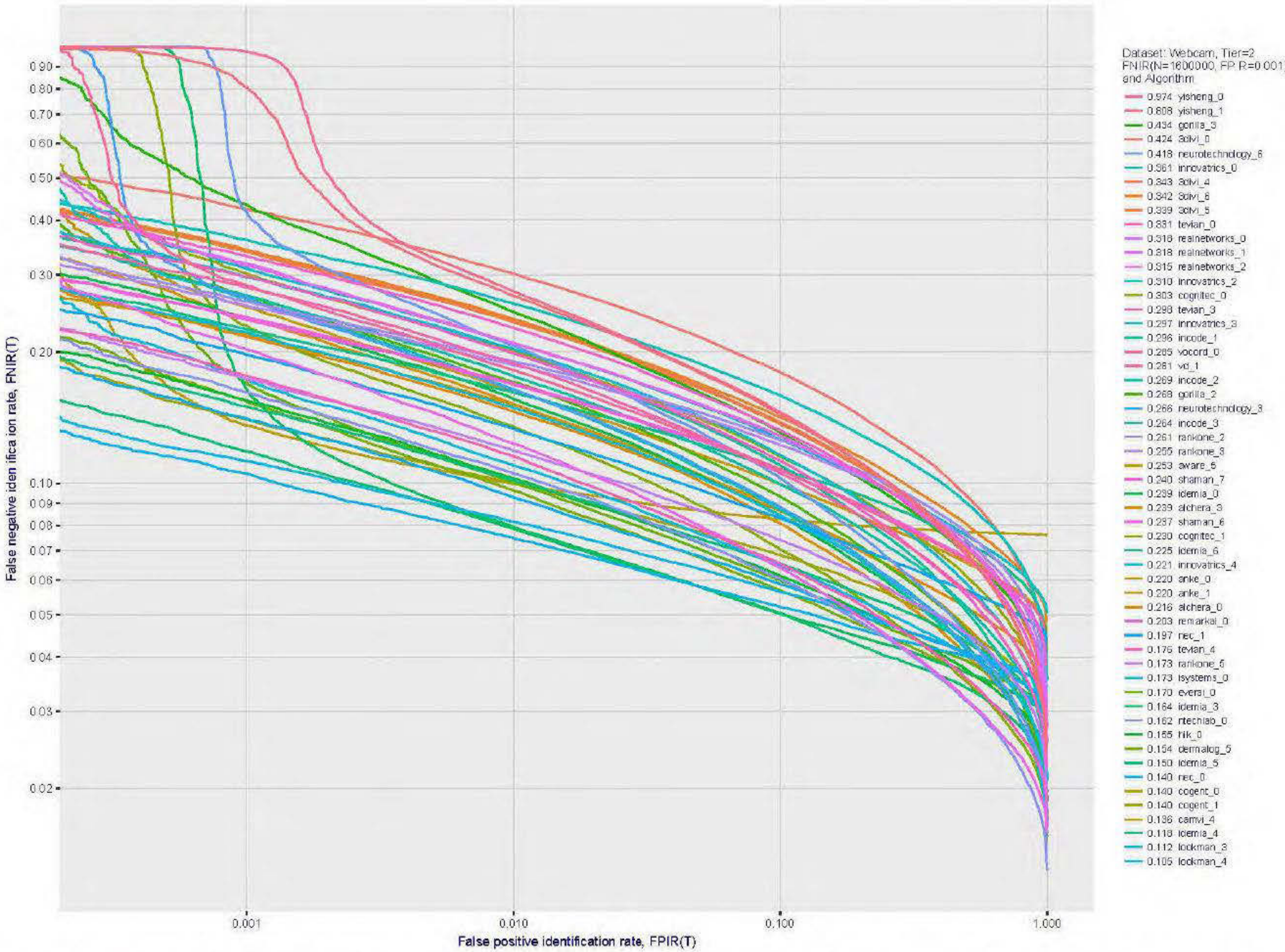


Figure 98: **[Webcam Dataset] Identification miss rates vs. false positive rates.** *The results apply to cross-domain recognition in which webcams are searched against enrolled mugshots. The FNIR values are higher than those for mugshot-mugshot identification due to low image resolution, lighting and less constrained subject pose in webcam images - see Figure 4.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

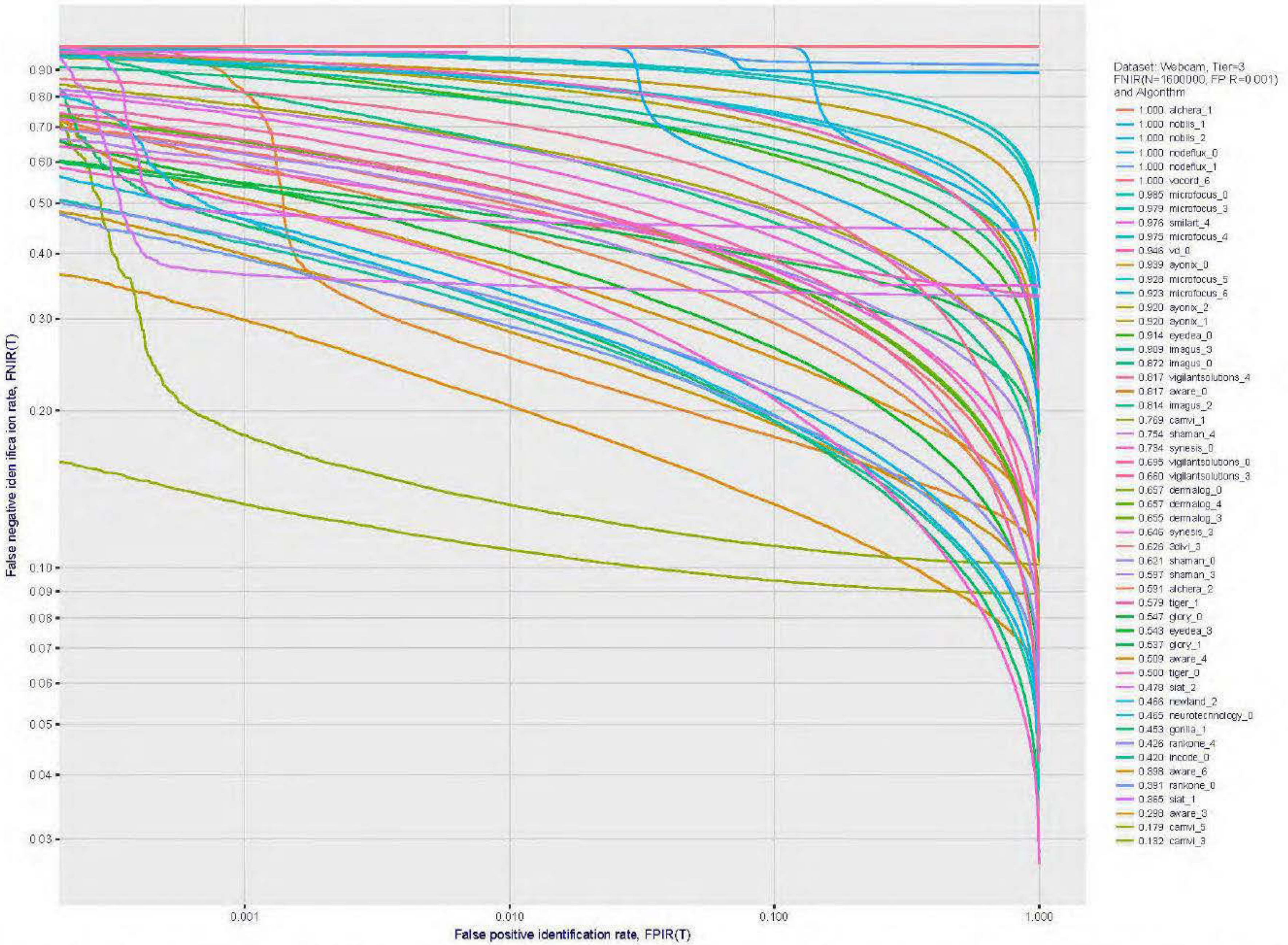


Figure 99: **[Webcam Dataset] Identification miss rates vs. false positive rates.** *The results apply to cross-domain recognition in which webcams are searched against enrolled mugshots. The FNIR values are higher than those for mugshot-mugshot identification due to low image resolution, lighting and less constrained subject pose in webcam images - see Figure 4.*

Appendix E Accuracy for profile-view to frontal recognition

Figures 100 - 102 gives accuracy results for searching 100 000 mated and 100 000 non-mated profile-view images against the same FRVT 2018 frontal enrollment dataset, N = 1 600 000, used in the main mugshot trials. This experiment corresponds to row-13 of Table 5. An example of profile-view image is given in Figure 5.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

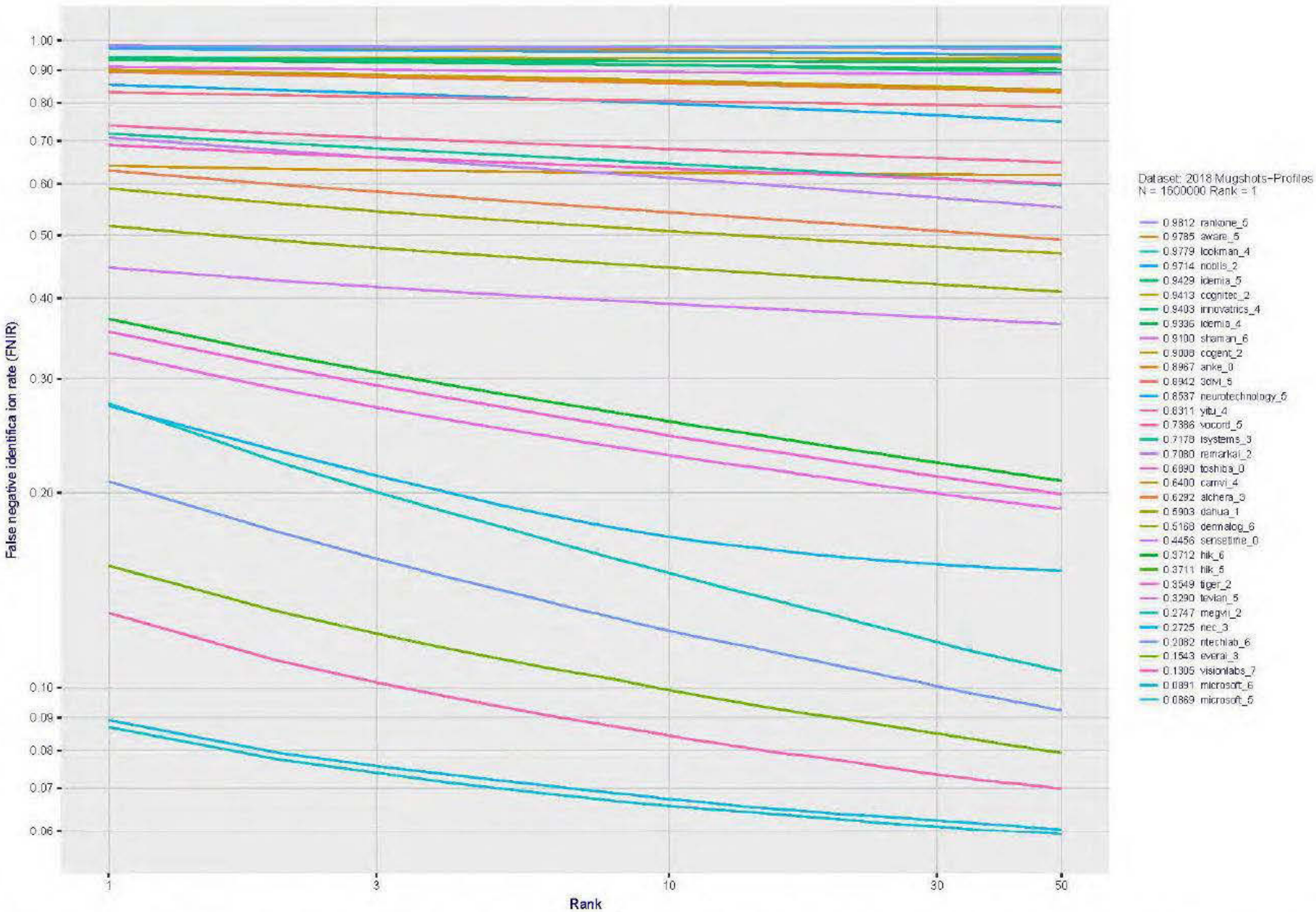


Figure 100: **[Mugshot and profile-view dataset] Rank-based accuracy.** *For some of the more accurate Phase 3 algorithms the figure plots error tradeoff characteristics for frontal and profile-view searches into an enrolled set of N = 1 600 000 frontal images. Note that some algorithms fail on profile-view images with FNIR → 1 - this evaluation did not ask developers to provide profile-view capability. Some algorithms, on the other hand, give FNIR approaching that for frontal-view searches using c. 2010 algorithms. The best result is that 91% of profile-view searches yield the correct mate at rank 1, and better than 94% in the top-50 candidates.*

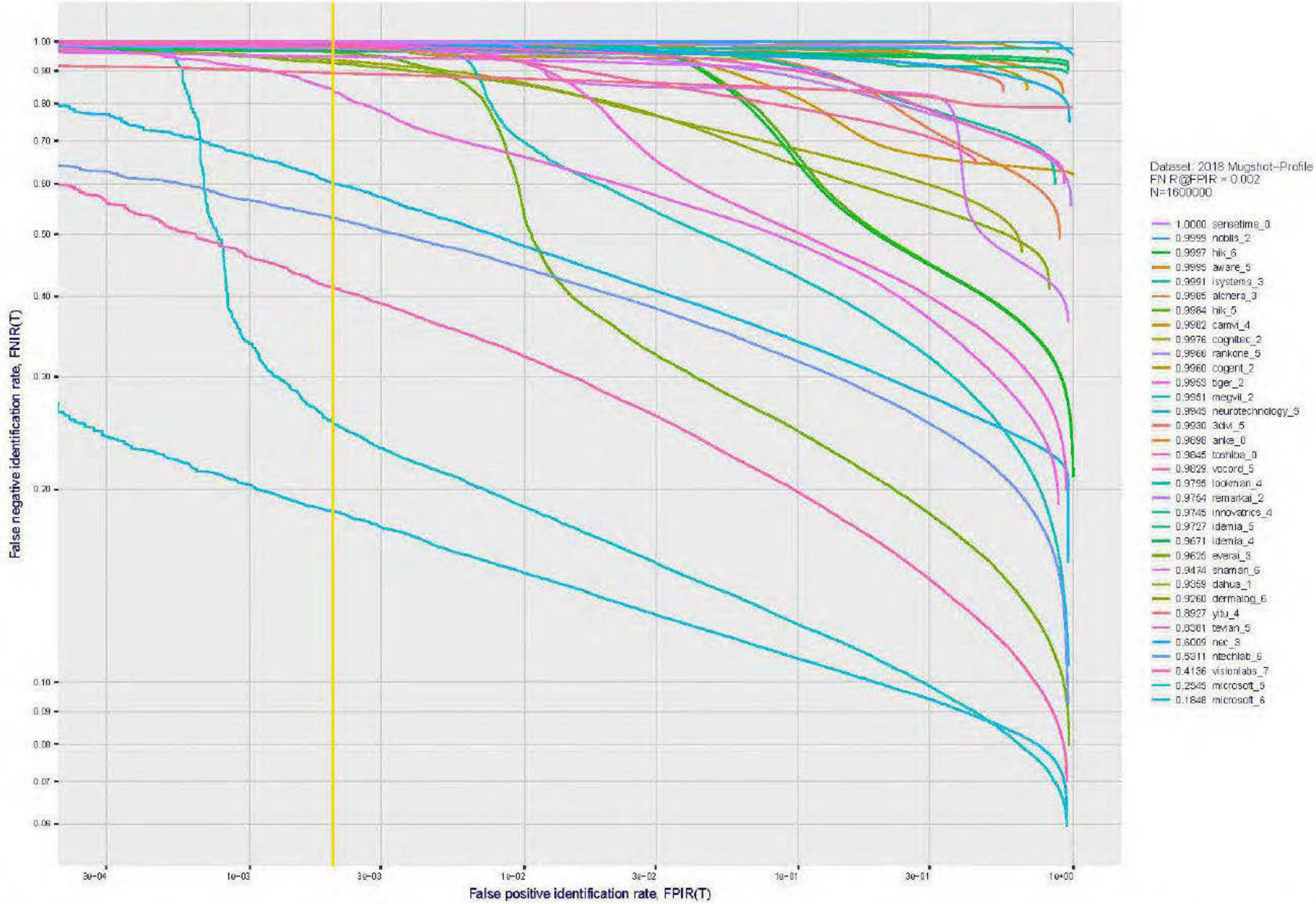


Figure 101: **[Mugshot and profile-view dataset] Threshold-based accuracy.** *For some of the more accurate Phase 3 algorithms the figure plots error tradeoff characteristics for frontal and profile-view searches into an enrolled set of N = 1 600 000 frontal images. Note that some algorithms fail on profile-view images with FNIR → 1 - this evaluation did not ask developers to provide profile-view capability. Some algorithms, on the other hand, give FNIR approaching that for frontal-view searches using c. 2010 algorithms.*

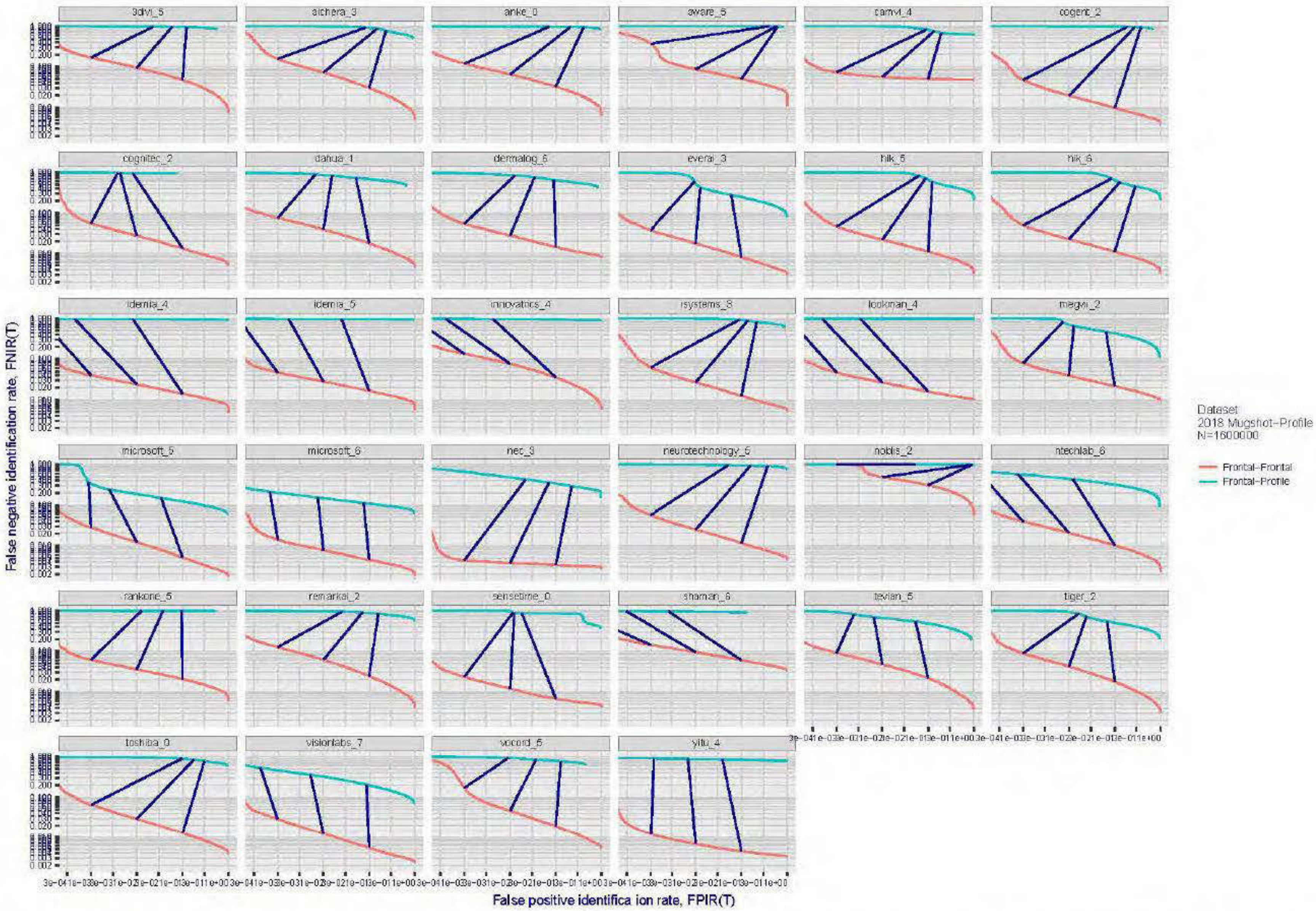


Figure 102: **[Mugshot and profile-view dataset] Speed-accuracy tradeoff**. *For some of the more accurate Phase 3 algorithms the figure plots error tradeoff characteristics for frontal and profile-view searches into an enrolled set of N = 1 600 000 frontal images. Some algorithms fail on profile-view images with FNIR → 1 - this evaluation did not ask developers to provide profile-view capability. Some algorithms, on the other hand, give FNIR approaching that for frontal-view searches using c. 2010 algorithms. Blue lines connect points of equal threshold from which it is evident that some algorithms would give markedly higher false positive outcomes if profile-view images were searched in a system configured for frontal searches. This would be a vulnerability in an access control system.*

Appendix F Accuracy when identifying wild images

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

2019/09/11	FNIR(N, R, T) =	False neg. identification rate	N = Num. enrolled subjects	T = Threshold	T = 0 → Investigation
17:24:52	FPIR(N, T) =	False pos. identification rate	R = Num. candidates examined		T > 0 → Identification

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

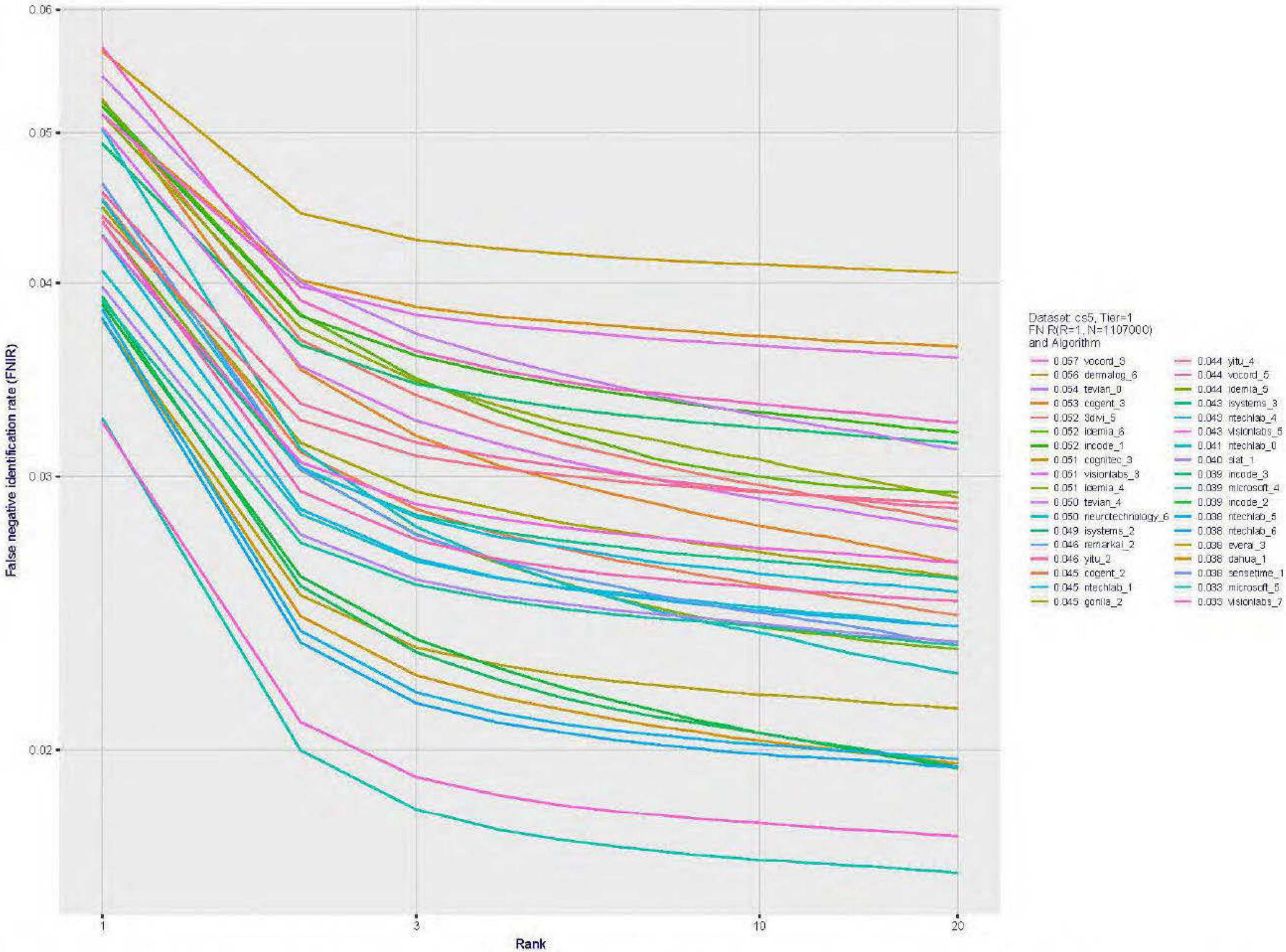


Figure 103: **[Wild Dataset] Identification miss rates vs. rank.** *For the wild dataset, the figure shows false negative identification rates (FNIR) vs. rank when the threshold is set to zero. This metric is relevant to human reviewers who will traverse candidate lists checking whether any of the returned identities match to the search imagery. Specifically, wild images were searched against 1.1 million individuals enrolled with wild images as well.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

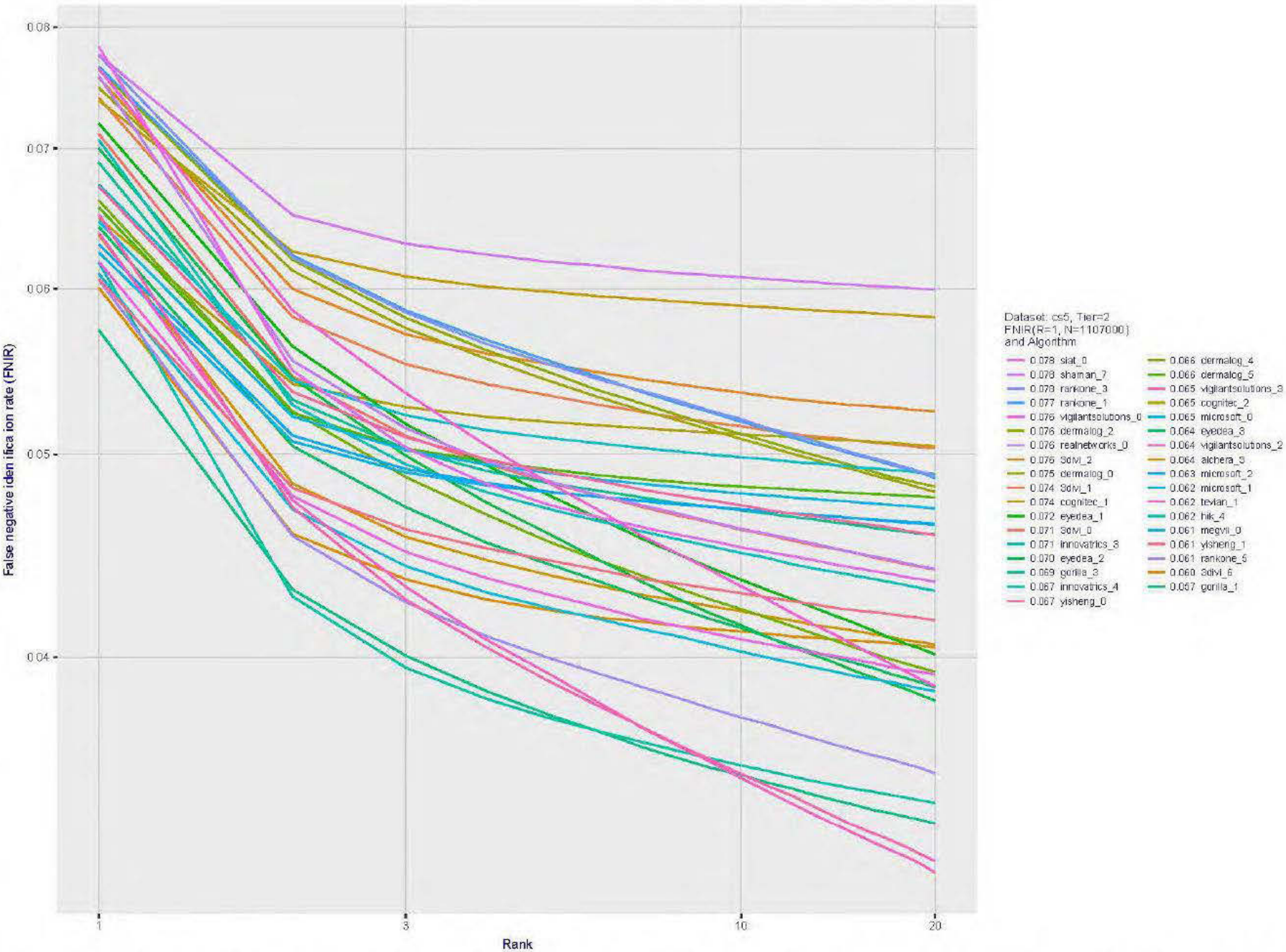


Figure 104: **[Wild Dataset] Identification miss rates vs. rank.** *For the wild dataset, the figure shows false negative identification rates (FNIR) vs. rank when the threshold is set to zero. This metric is relevant to human reviewers who will traverse candidate lists checking whether any of the returned identities match to the search imagery. Specifically, wild images were searched against 1.1 million individuals enrolled with wild images as well.*

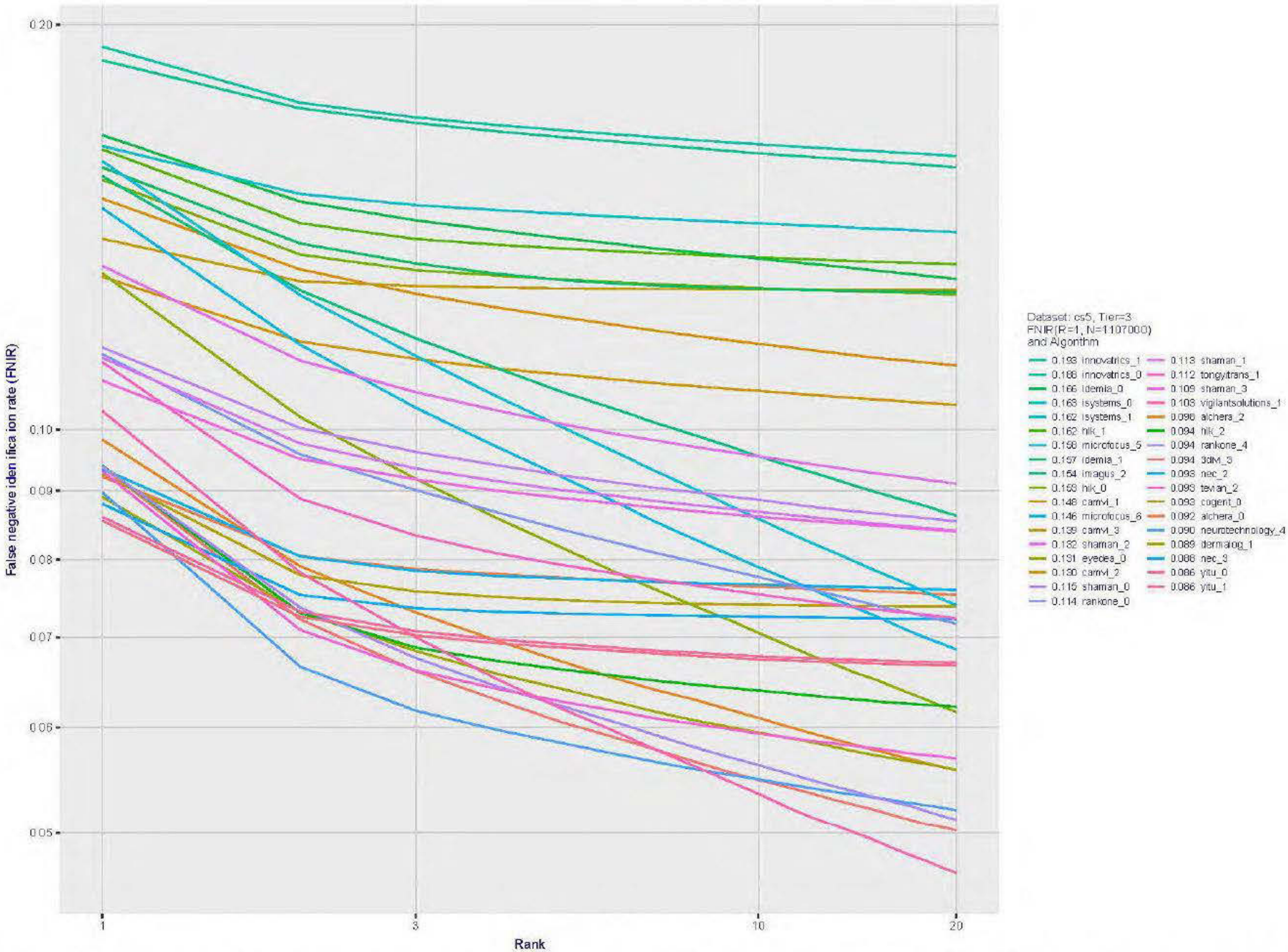


Figure 105: **[Wild Dataset] Identification miss rates vs. rank.** *For the wild dataset, the figure shows false negative identification rates (FNIR) vs. rank when the threshold is set to zero. This metric is relevant to human reviewers who will traverse candidate lists checking whether any of the returned identities match to the search imagery. Specifically, wild images were searched against 1.1 million individuals enrolled with wild images as well.*

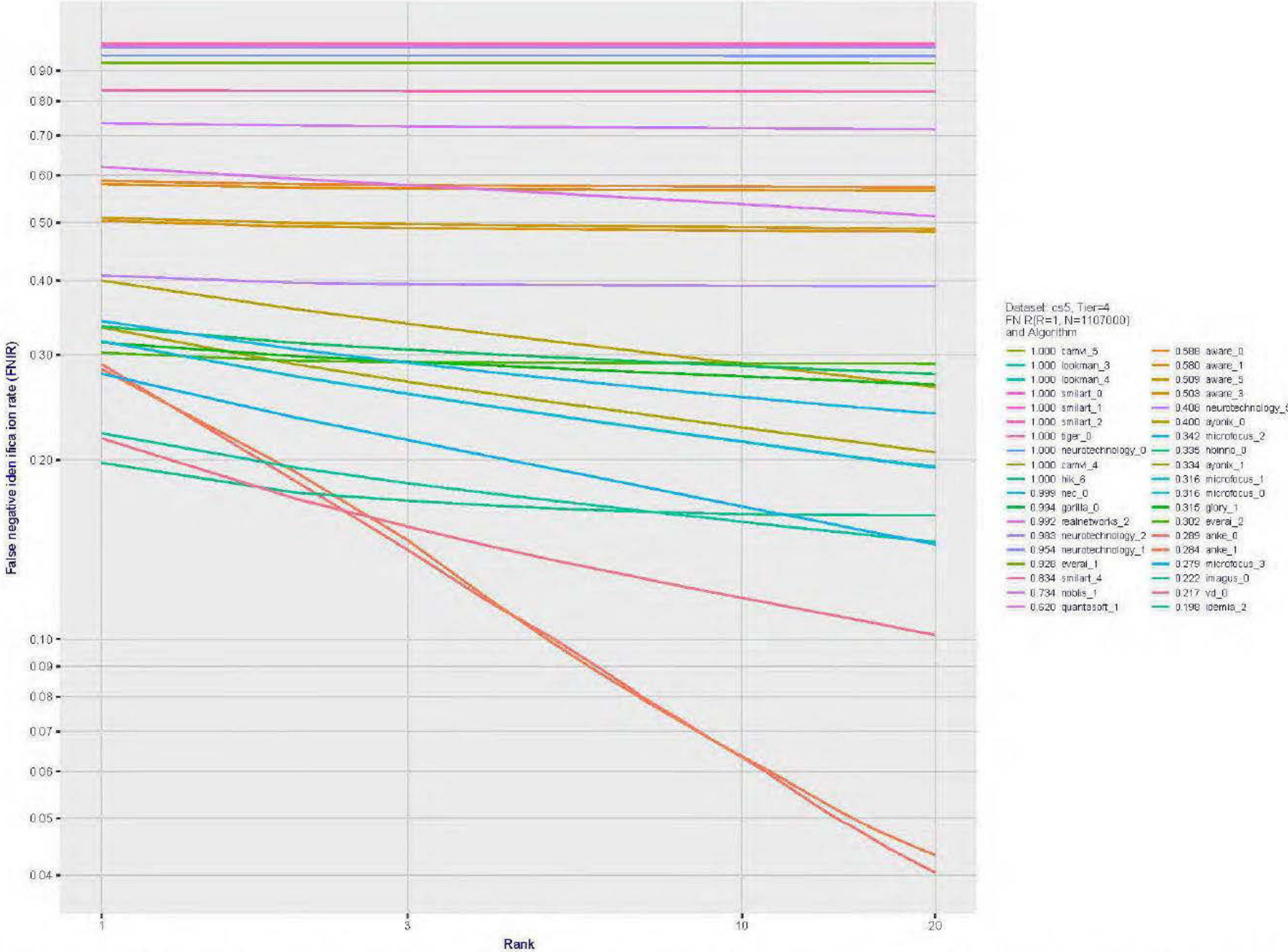


Figure 106: **[Wild Dataset] Identification miss rates vs. rank.** *For the wild dataset, the figure shows false negative identification rates (FNIR) vs. rank when the threshold is set to zero. This metric is relevant to human reviewers who will traverse candidate lists checking whether any of the returned identities match to the search imagery. Specifically, wild images were searched against 1.1 million individuals enrolled with wild images as well.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

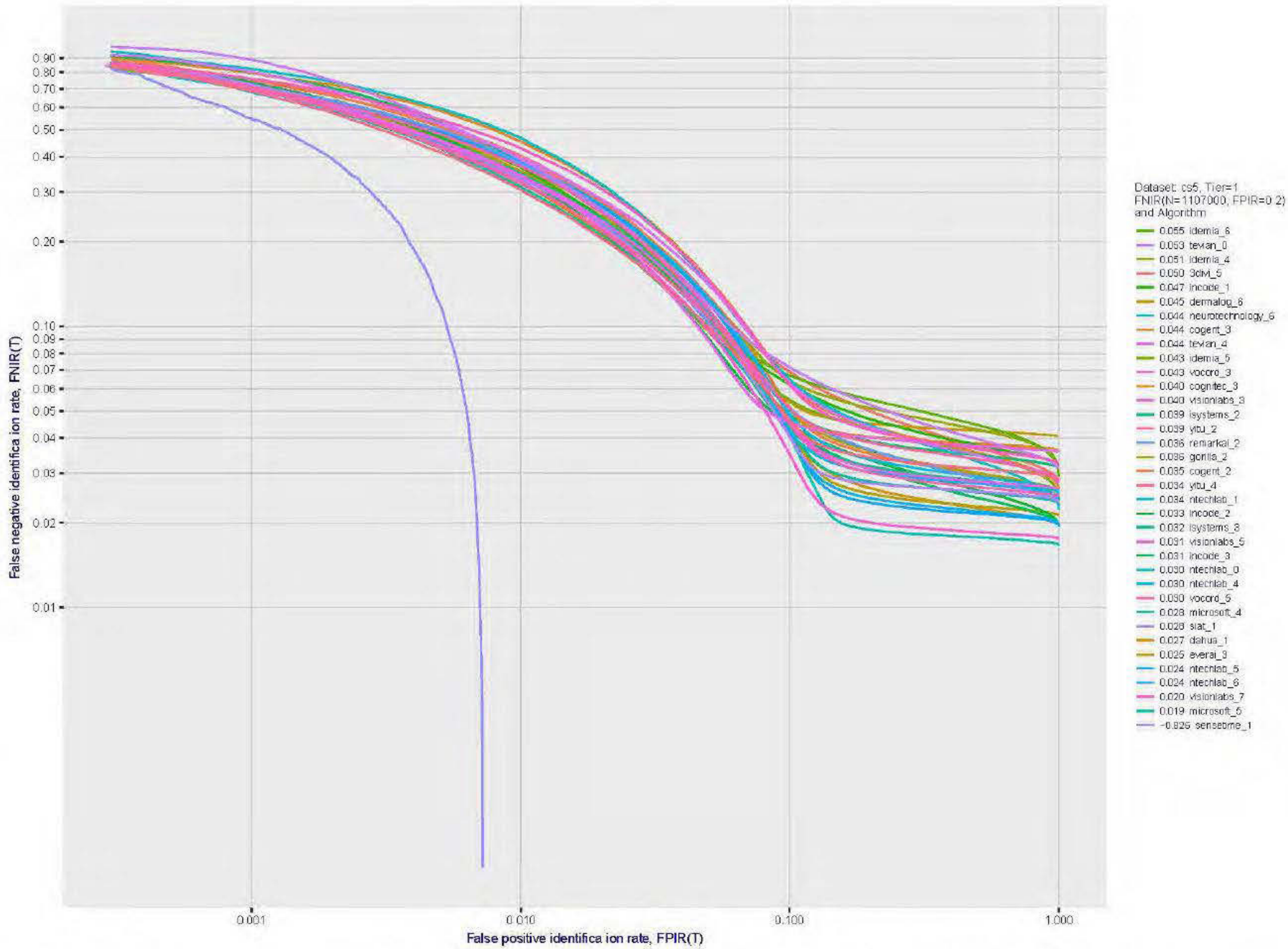


Figure 107: **[Wild Dataset] Identification miss rates vs. false positive rates.** *The figure shows accuracy of algorithms on wild images searched against wild images of 1.1 million individuals enrolled into a gallery. On the vertical axis is miss rate FNIR(N, T, L) with N = 1 107 000, as a function of false positive identification FPIR(N, T). The rapid increase in FNIR below FPIR = 0.1 suggests that some background identities in the gallery are actually present in the non-mated search sets. This issue will be addressed in the 2019 revision of this report.*

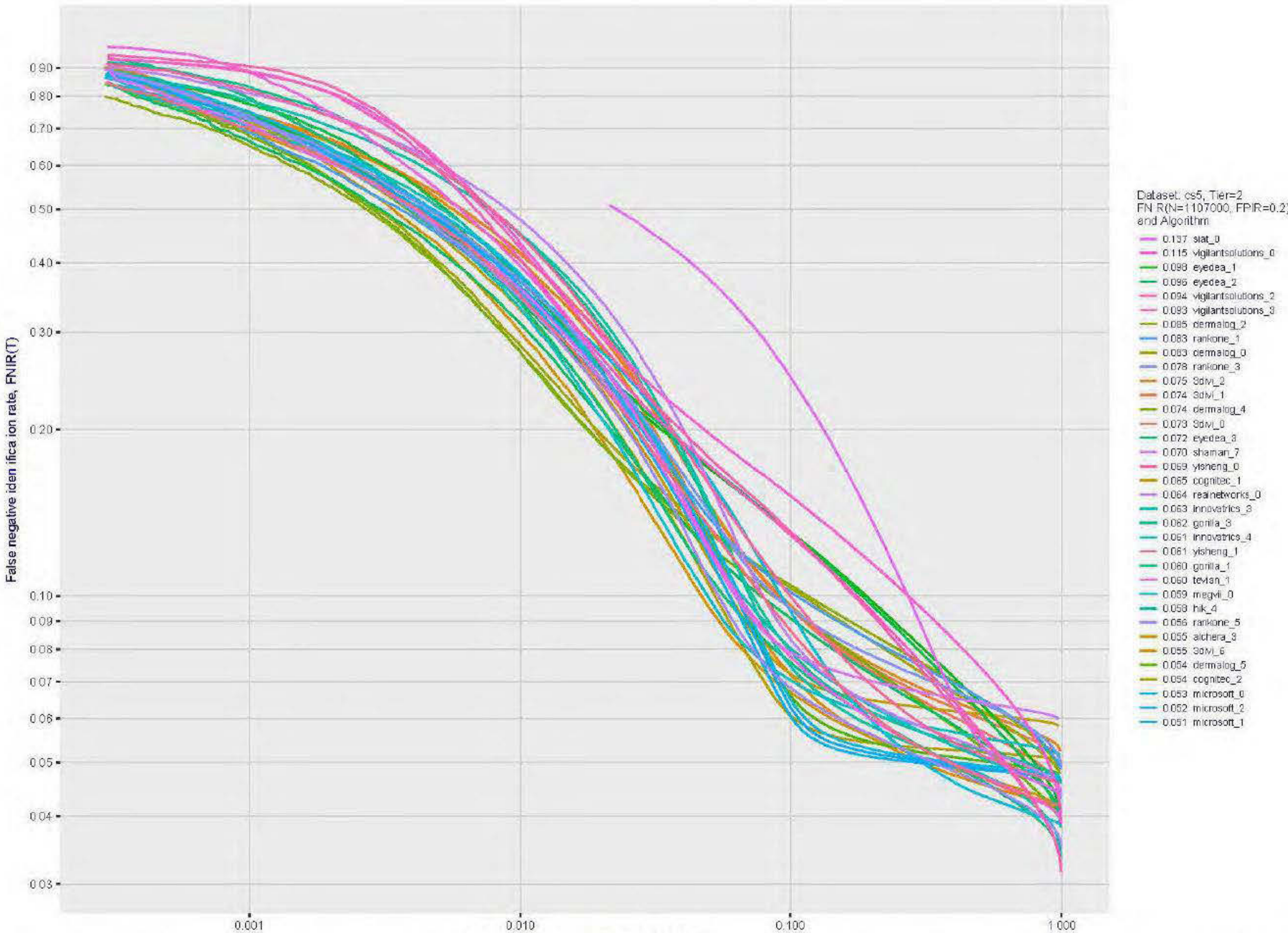


Figure 108: **[Wild Dataset] Identification miss rates vs. false positive rates**. *The figure shows accuracy of algorithms on wild images searched against wild images of 1.1 million individuals enrolled into a gallery. On the vertical axis is miss rate FNIR(N, T, L) with N = 1 107 000, as a function of false positive identification FPIR(N, T). The rapid increase in FNIR below FPIR = 0.1 suggests that some background identities in the gallery are actually present in the non-mated search sets. This issue will be addressed in the 2019 revision of this report.*

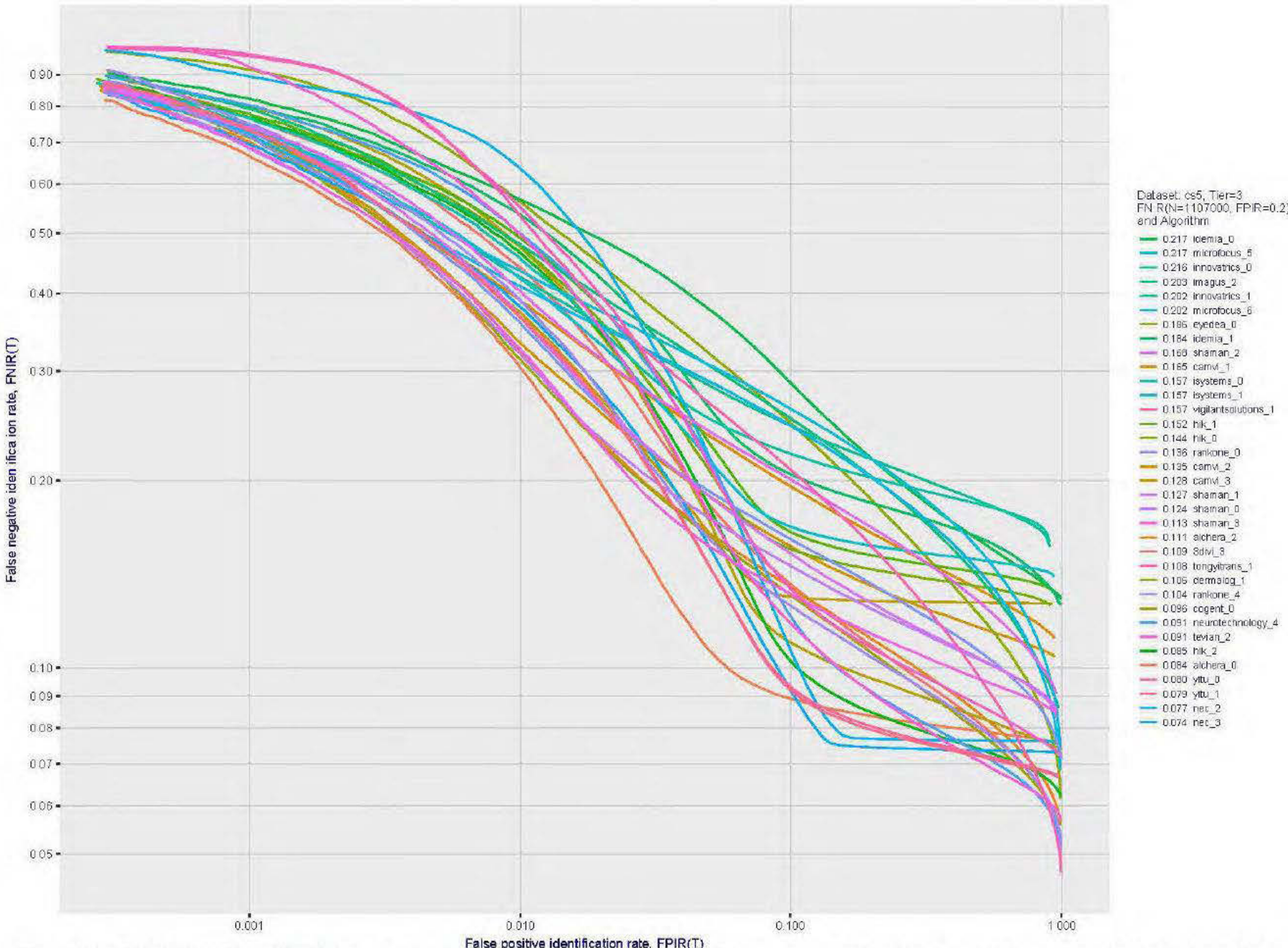


Figure 109: **[Wild Dataset] Identification miss rates vs. false positive rates.** *The figure shows accuracy of algorithms on wild images searched against wild images of 1.1 million individuals enrolled into a gallery. On the vertical axis is miss rate FNIR(N, T, L) with N = 1 107 000, as a function of false positive identification FPIR(N, T). The rapid increase in FNIR below FPIR = 0.1 suggests that some background identities in the gallery are actually present in the non-mated search sets. This issue will be addressed in the 2019 revision of this report.*

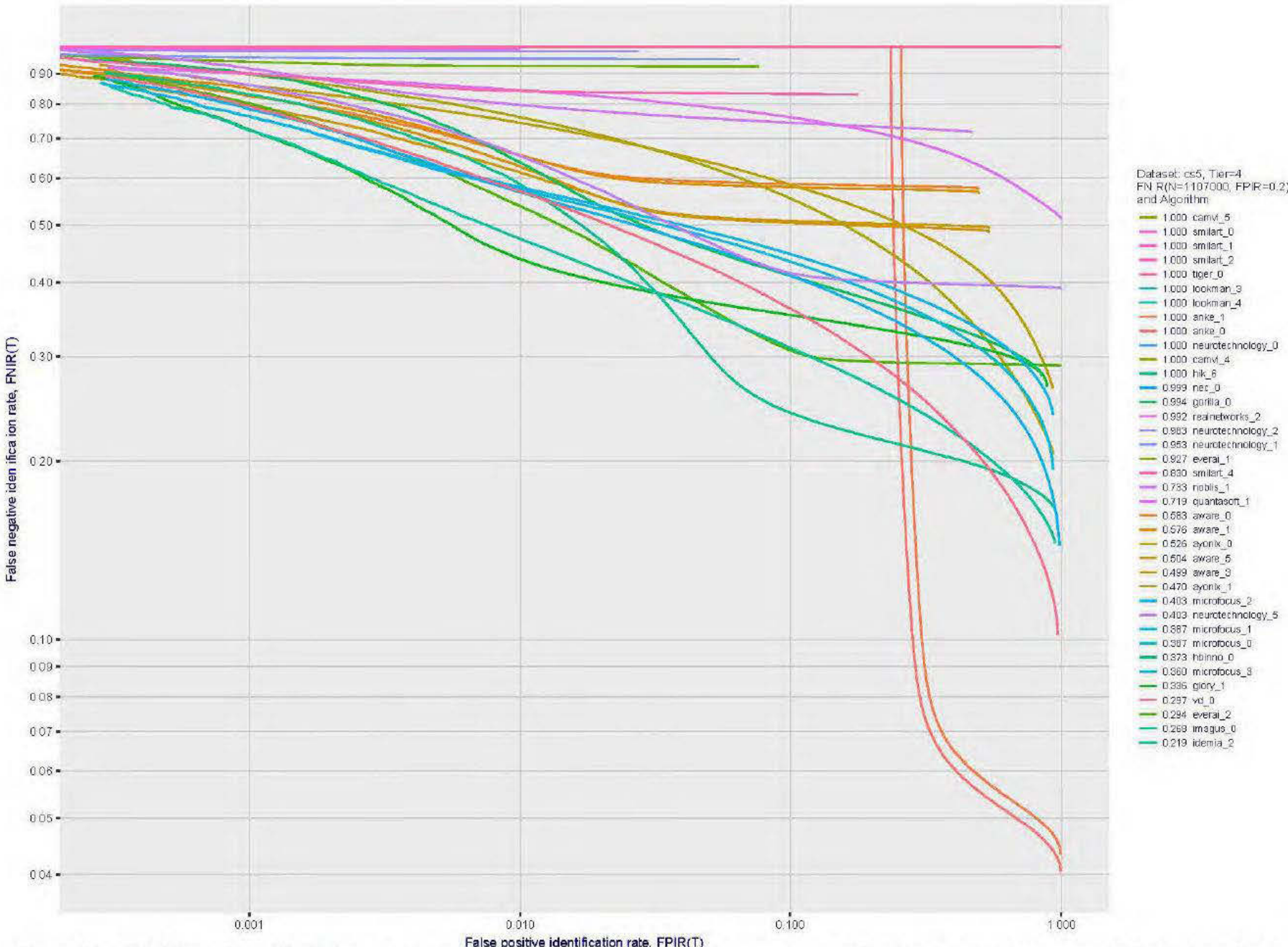


Figure 110: **[Wild Dataset] Identification miss rates vs. false positive rates.** *The figure shows accuracy of algorithms on wild images searched against wild images of 1.1 million individuals enrolled into a gallery. On the vertical axis is miss rate FNIR(N, T, L) with N = 1 107 000, as a function of false positive identification FPIR(N, T). The rapid increase in FNIR below FPIR = 0.1 suggests that some background identities in the gallery are actually present in the non-mated search sets. This issue will be addressed in the 2019 revision of this report.*

Appendix G Search duration

As in and prior tests, this section documents search speeds spanning three orders of magnitude. In applications where search volumes are high enough, this will have implications for hardware requirements especially for large N or when search duration is appreciably larger than the time it takes to prepare a template from the search image(s). Further, given very large (and growing) operational databases, the scalability of algorithms is important. It has been reported previously [8] that search duration can scale sublinearly with enrolled population size N. Further there has been considerable recent research on indexing, exact [13] and approximate nearest neighbor search [1,13] and fast-search [14,16].

Figure 111 charts the search duration measurements presented earlier in Tables 6 - 9.

- ▷ Most algorithms scale linearly. For those in that category, there is a wide range in speed with search durations ranging from 82 milliseconds for a 12 million gallery (for NEC-3) to more than 40 seconds (for Yitu-3, Toshiba-2) and even higher for less accurate algorithms.
- ▷ Some developers (Camvi, Dermalog, EverAI, Innovatrics, and Visionlabs) provide algorithms whose template search durations grow logarithmically i.e. approximately $T(N) = a \log N$ with the constant a varying between implementations. In the figure this model is fit using the point $T(1) = 0$, and $T(640\,000)$. This very sublinear behaviour affords extremely fast search times in very large galleries. One caveat for the sublinear algorithms is that the fast-search data structures require considerable computation time - on the order of hours - for N in the millions, and this scales mildy super-linearly, i.e. $O(N^b), b > 1$. There are exceptions: the Camvi algorithms take minutes; and Innvovatrics' scale sublinearly.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

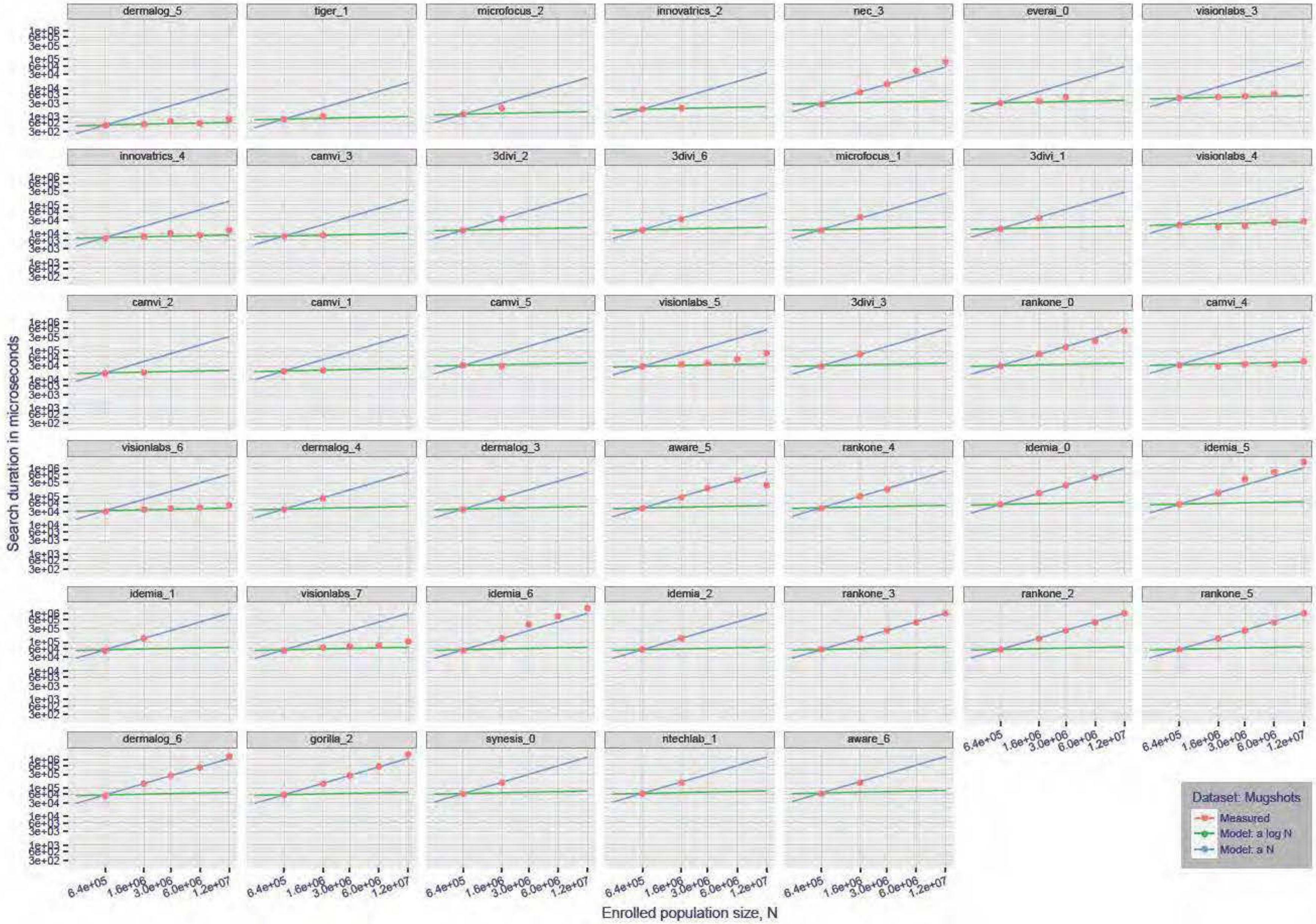


Figure 111: **[Mugshot Dataset] Search duration vs. enrolled population size**. *In red are the actual point durations measured on a single c. 2016 core. The blue shows linear growth from $N = 640\,000$. The green line shows logathmic growth from that point to $N = 1\,600\,000$. Note the sublinear growth from algorithms from Camvi, Dermalog, EverAI, Innovatrics, and Visionlabs. The tiger_1 algorithm is also sublinear, but inaccurate and inoperable at $N \geq 3000000$. This capability sometimes comes at the additional expense of converting a linear gallery data structure into whatever fast-search data structure is used. Note that search times are sometimes dominated by the template generation times shown in Table 16.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

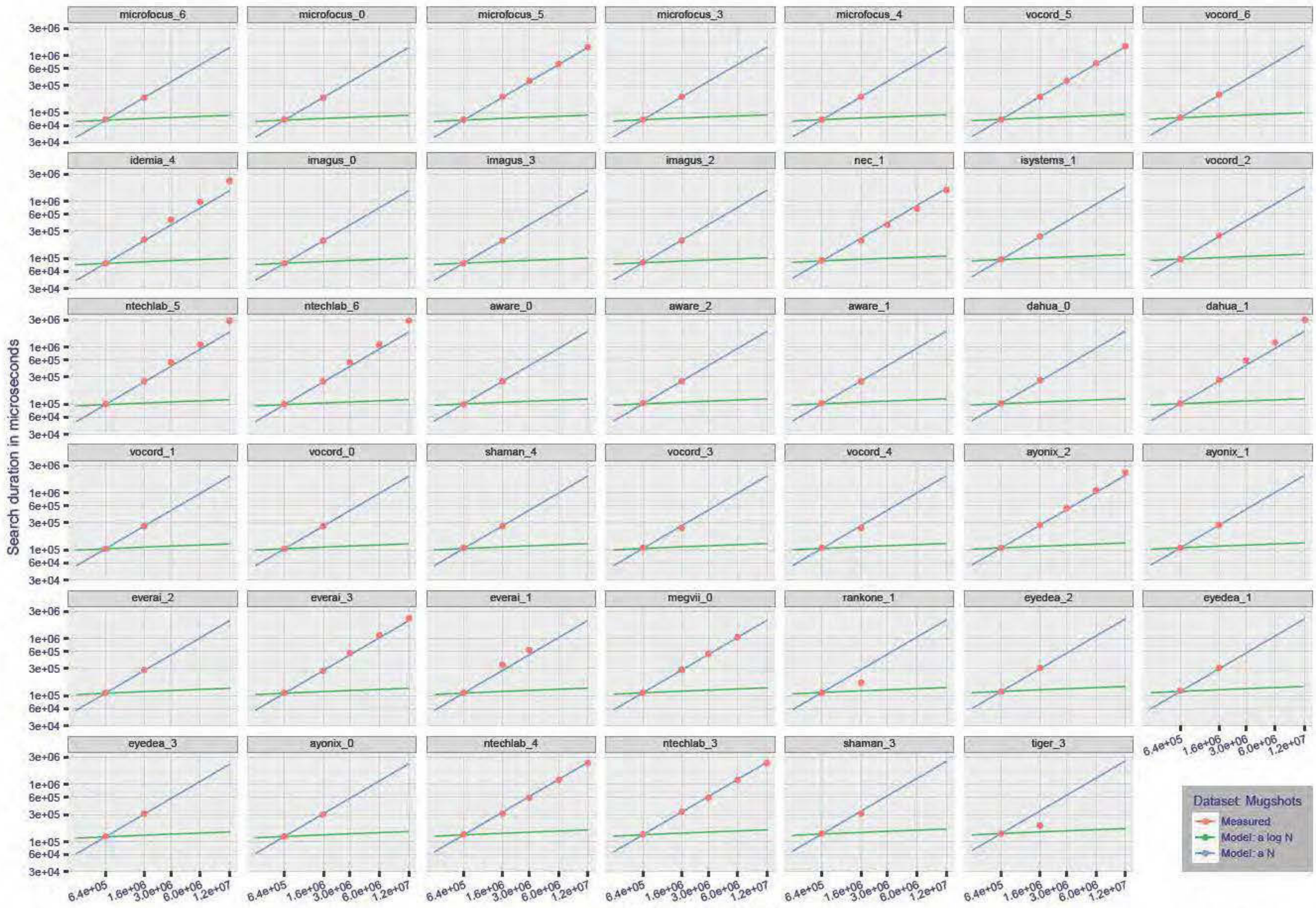


Figure 112: **[Mugshot Dataset] Search duration vs. enrolled population size.** *In red are the actual point durations measured on a single c. 2016 core. The blue shows linear growth from $N = 640\,000$. The green line shows logathmic growth from that point to $N = 1\,600\,000$. Note the sublinear growth from algorithms from Camvi, Dermalog, EverAI, Innovatrics, and Visionlabs. The tiger_1 algorithm is also sublinear, but inaccurate and inoperable at $N \geq 3000000$. This capability sometimes comes at the additional expense of converting a linear gallery data structure into whatever fast-search data structure is used. Note that search times are sometimes dominated by the template generation times shown in Table 16.*

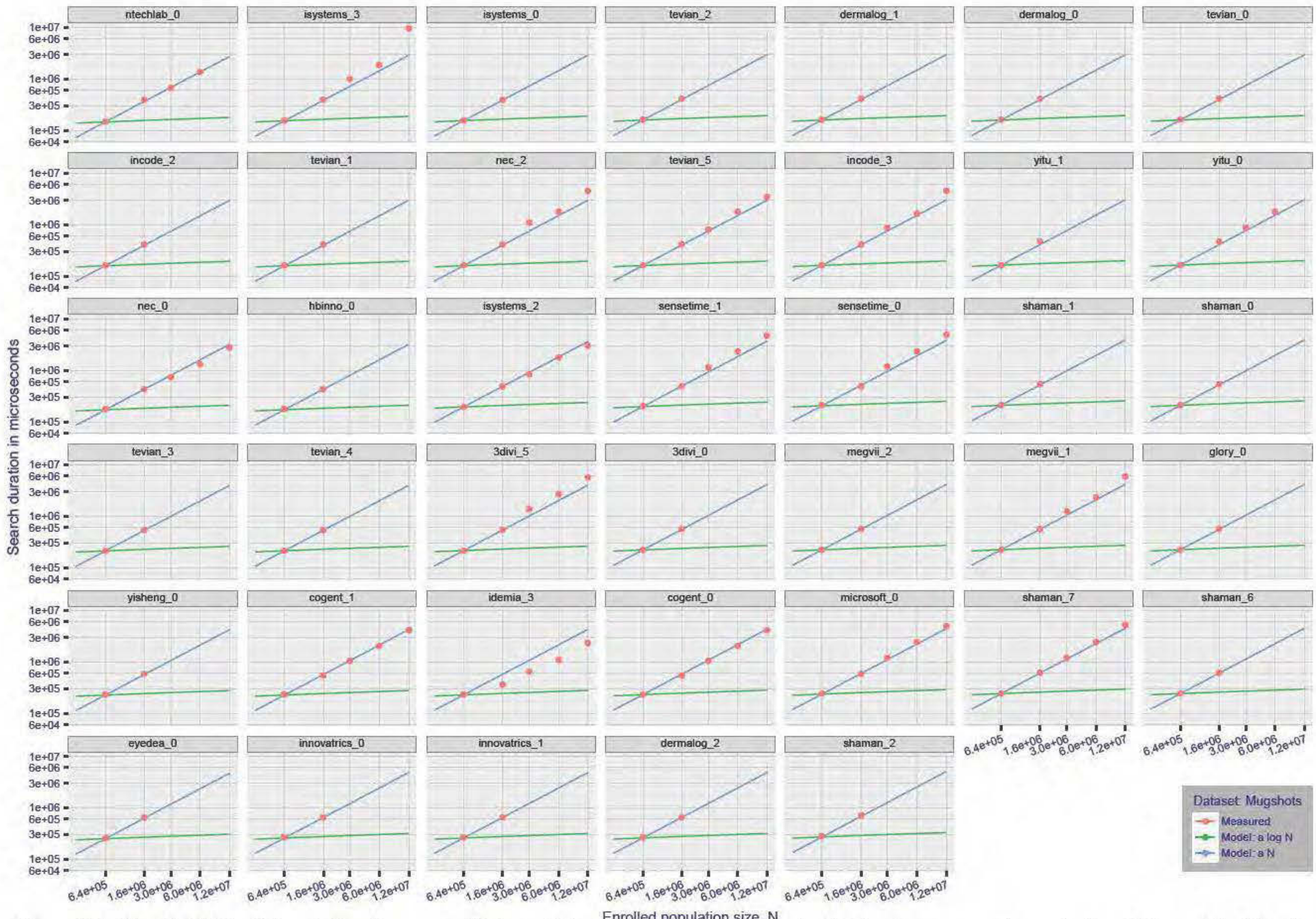


Figure 113: **[Mugshot Dataset] Search duration vs. enrolled population size.** *In red are the actual point durations measured on a single c. 2016 core. The blue shows linear growth from $N = 640\,000$. The green line shows logathmic growth from that point to $N = 1\,600\,000$. Note the sublinear growth from algorithms from Camvi, Dermalog, EverAI, Innovatrics, and Visionlabs. The tiger_1 algorithm is also sublinear, but inaccurate and inoperable at $N \geq 3000000$. This capability sometimes comes at the additional expense of converting a linear gallery data structure into whatever fast-search data structure is used. Note that search times are sometimes dominated by the template generation times shown in Table 16.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

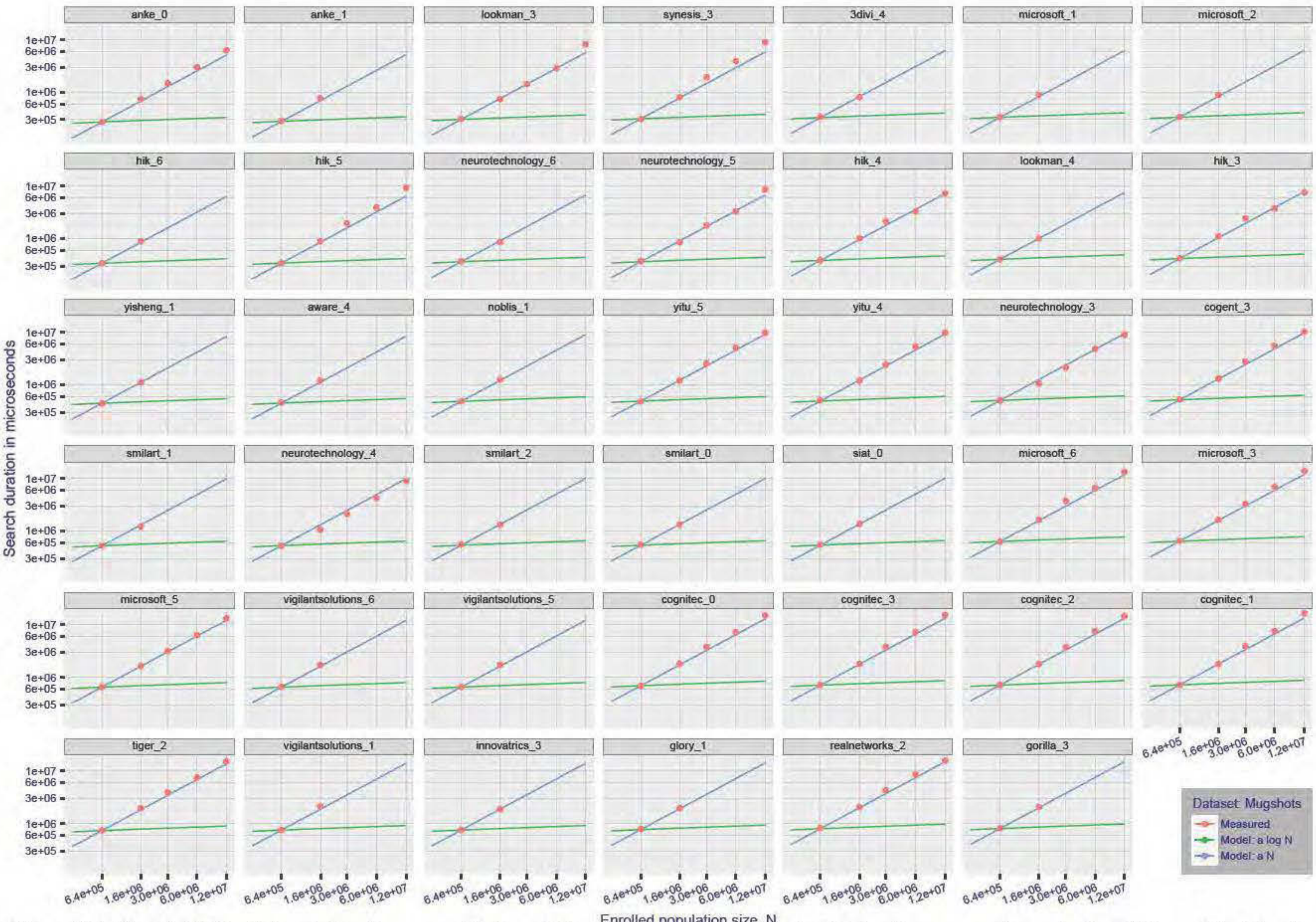


Figure 114: **[Mugshot Dataset] Search duration vs. enrolled population size.** *In red are the actual point durations measured on a single c. 2016 core. The blue shows linear growth from $N = 640\,000$. The green line shows logathmic growth from that point to $N = 1\,600\,000$. Note the sublinear growth from algorithms from Camvi, Dermalog, EverAI, Innovatrics, and Visionlabs. The tiger_1 algorithm is also sublinear, but inaccurate and inoperable at $N \geq 3000000$. This capability sometimes comes at the additional expense of converting a linear gallery data structure into whatever fast-search data structure is used. Note that search times are sometimes dominated by the template generation times shown in Table 16.*

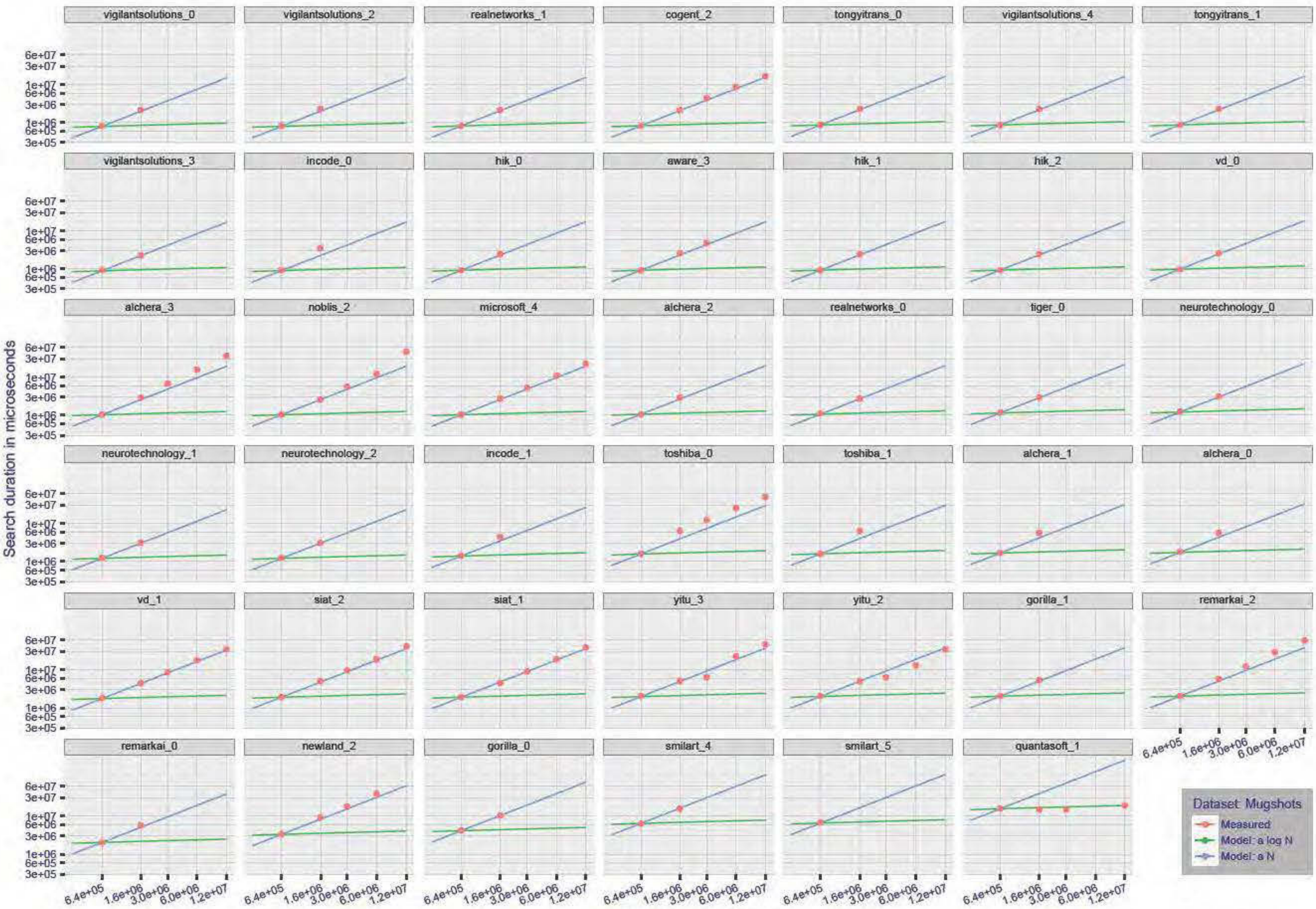


Figure 115: **[Mugshot Dataset] Search duration vs. enrolled population size.** *In red are the actual point durations measured on a single c. 2016 core. The blue shows linear growth from $N = 640\,000$. The green line shows logathmic growth from that point to $N = 1\,600\,000$. Note the sublinear growth from algorithms from Camvi, Dermalog, EverAI, Innovatrics, and Visionlabs. The tiger_1 algorithm is also sublinear, but inaccurate and inoperable at $N \geq 3000000$. This capability sometimes comes at the additional expense of converting a linear gallery data structure into whatever fast-search data structure is used. Note that search times are sometimes dominated by the template generation times shown in Table 16.*

Appendix H Gallery Insertion Timing

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

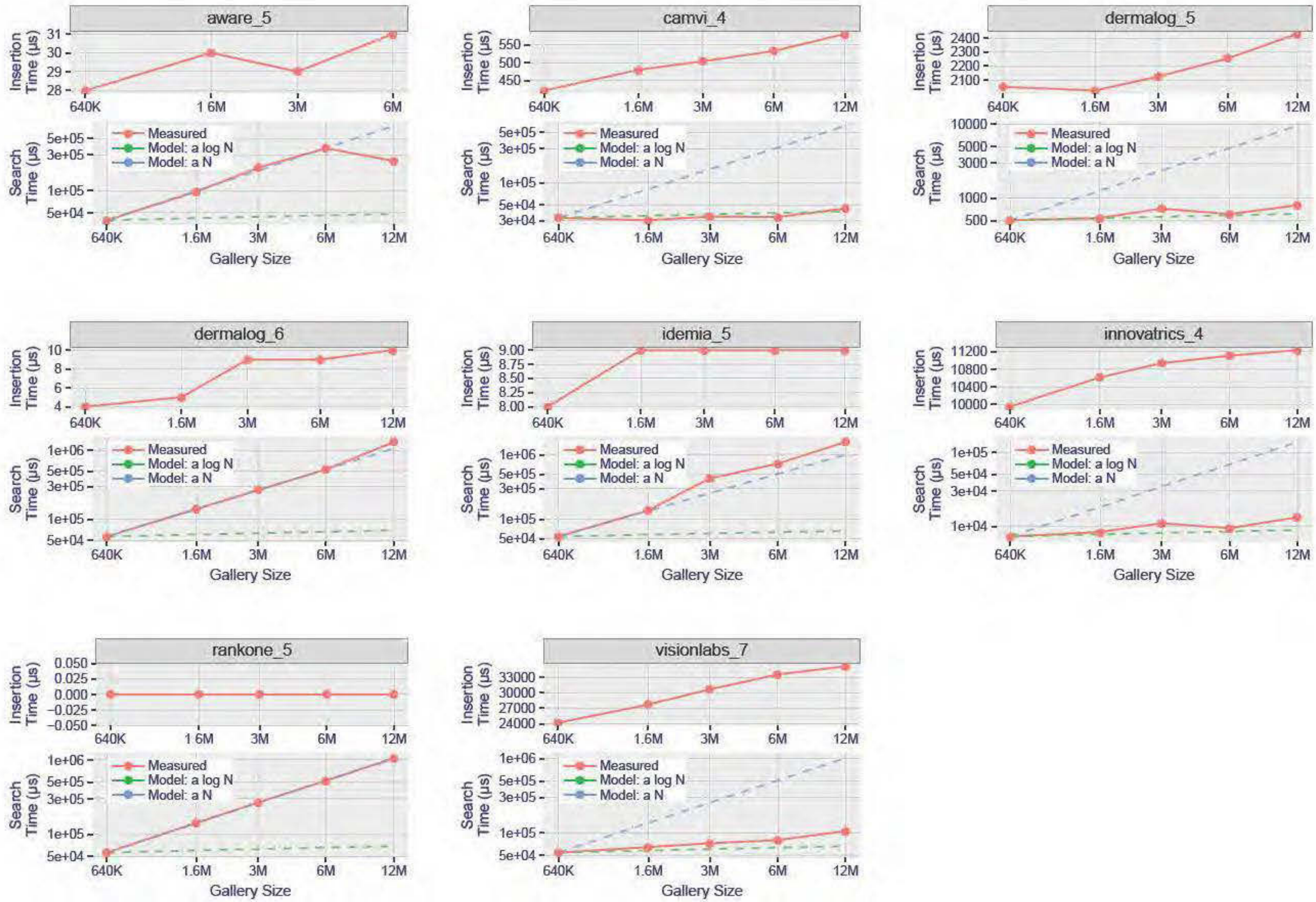


Figure 116: **[Mugshot Dataset] Gallery insertion duration vs. enrolled population size.** *This chart plots the time it takes to insert a single template into a finalized gallery, illustrated over increasing gallery sizes. For reference, search times on finalized galleries of corresponding sizes are plotted right underneath. Gallery insertion time plots were generated on algorithms that 1) successfully implemented gallery insertion with no errors and 2) that were run on galleries with N up to 12 000 000. Generally, only the more accurate algorithms were run on galleries with N up to 12 000 000.*

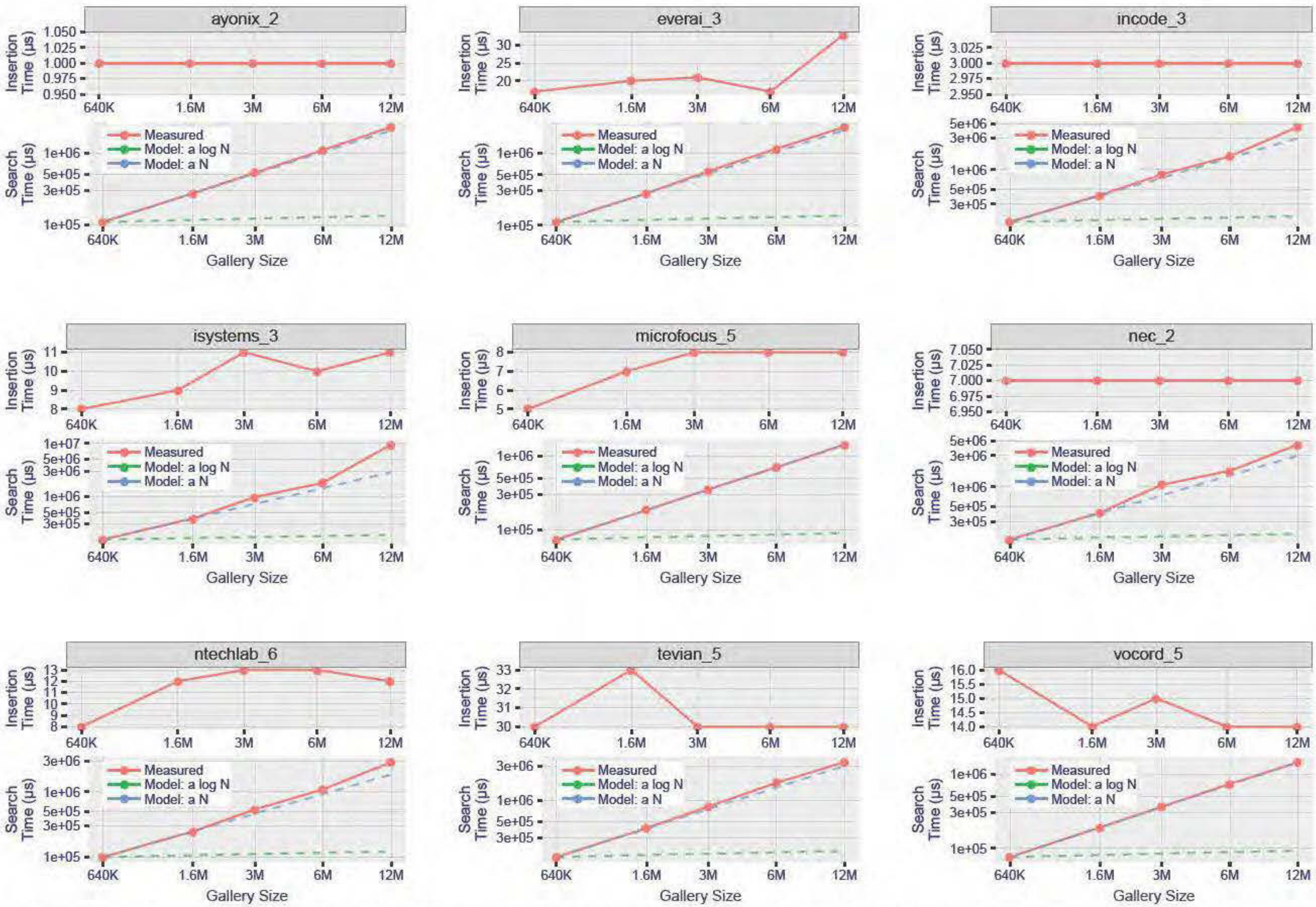


Figure 117: **[Mugshot Dataset] Gallery insertion duration vs. enrolled population size.** *This chart plots the time it takes to insert a single template into a finalized gallery, illustrated over increasing gallery sizes. For reference, search times on finalized galleries of corresponding sizes are plotted right underneath. Gallery insertion time plots were generated on algorithms that 1) successfully implemented gallery insertion with no errors and 2) that were run on galleries with N up to 12 000 000. Generally, only the more accurate algorithms were run on galleries with N up to 12 000 000.*

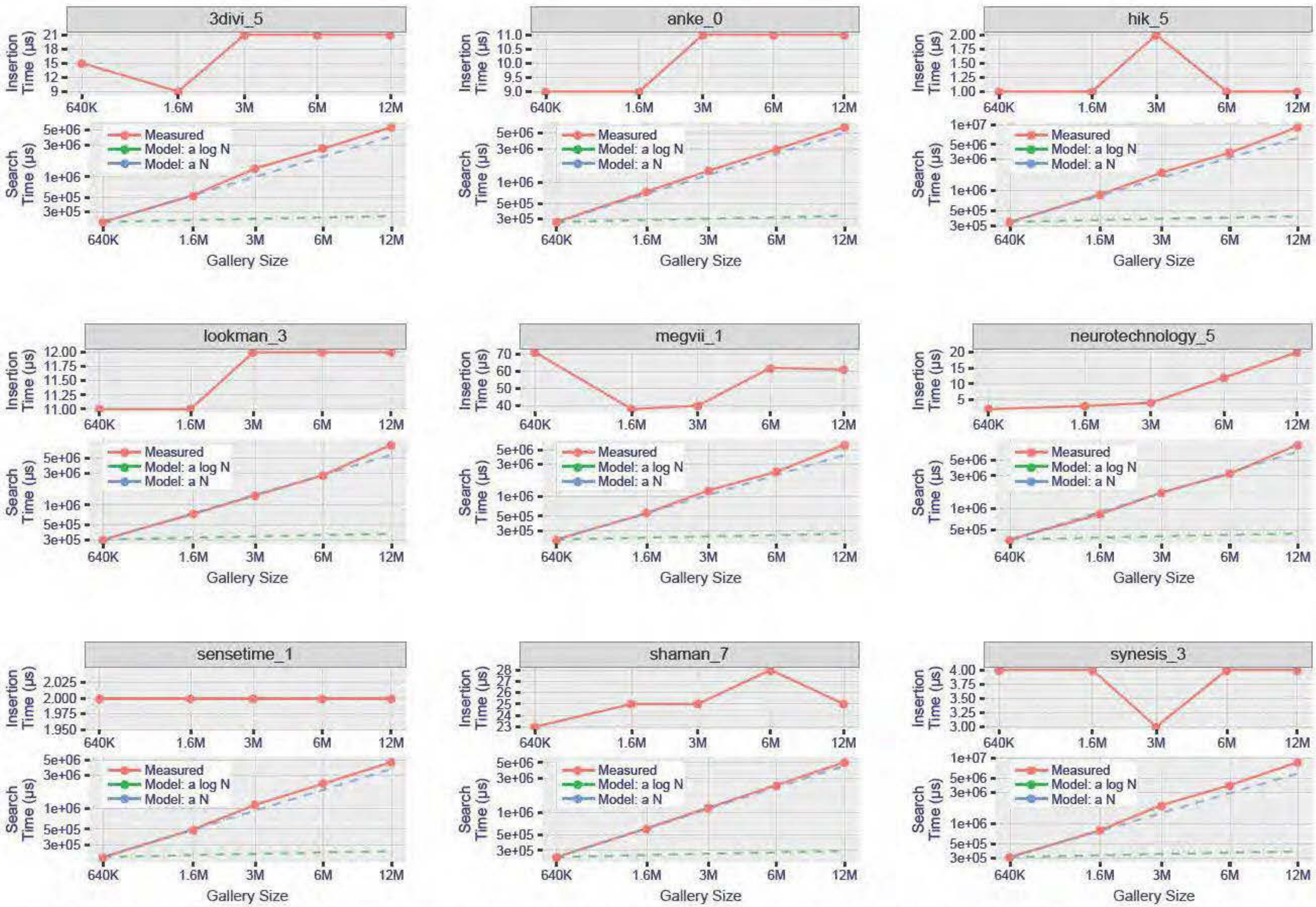


Figure 118: **[Mugshot Dataset] Gallery insertion duration vs. enrolled population size.** *This chart plots the time it takes to insert a single template into a finalized gallery, illustrated over increasing gallery sizes. For reference, search times on finalized galleries of corresponding sizes are plotted right underneath. Gallery insertion time plots were generated on algorithms that 1) successfully implemented gallery insertion with no errors and 2) that were run on galleries with N up to 12 000 000. Generally, only the more accurate algorithms were run on galleries with N up to 12 000 000.*

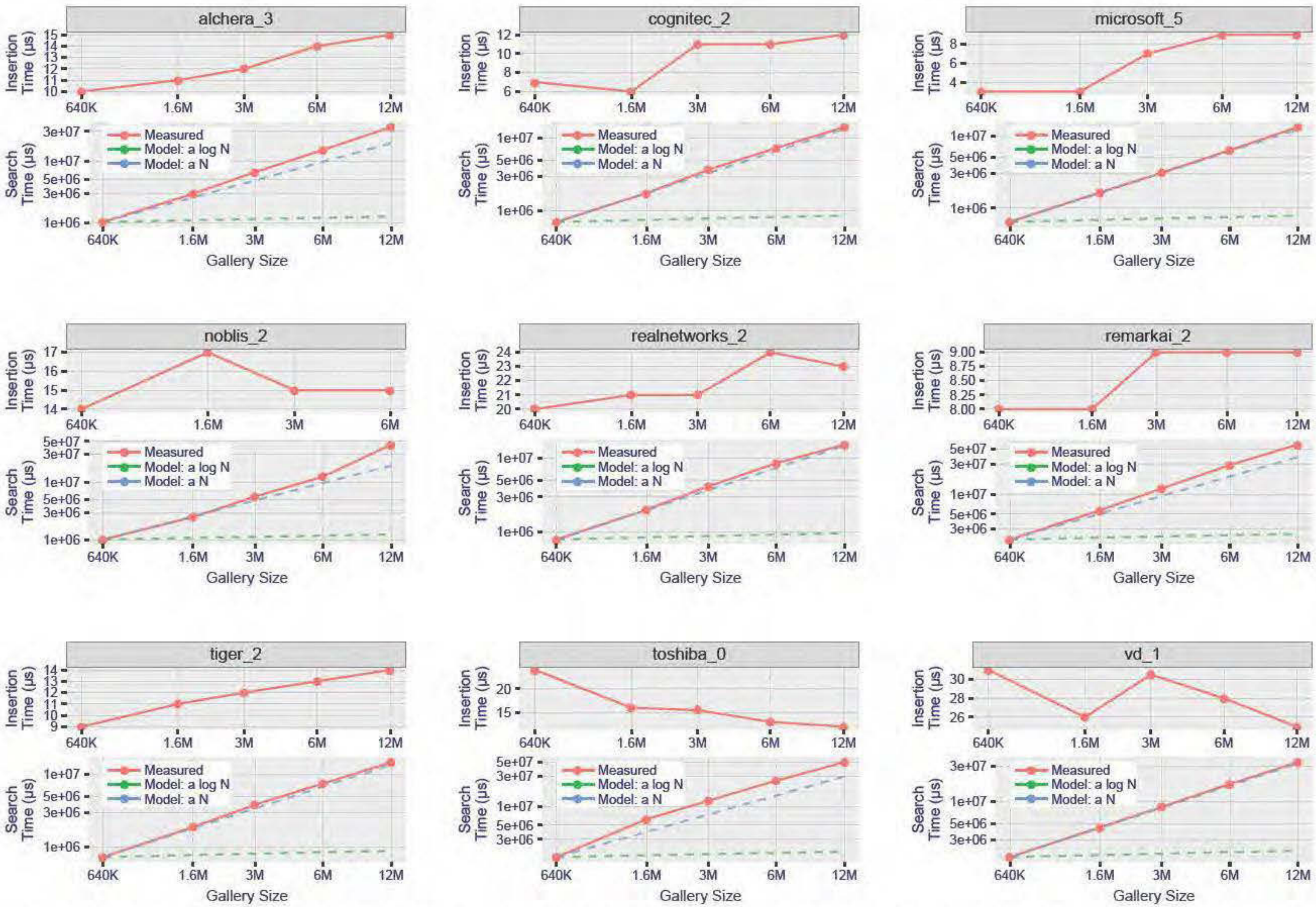


Figure 119: **[Mugshot Dataset] Gallery insertion duration vs. enrolled population size.** *This chart plots the time it takes to insert a single template into a finalized gallery, illustrated over increasing gallery sizes. For reference, search times on finalized galleries of corresponding sizes are plotted right underneath. Gallery insertion time plots were generated on algorithms that 1) successfully implemented gallery insertion with no errors and 2) that were run on galleries with N up to 12 000 000. Generally, only the more accurate algorithms were run on galleries with N up to 12 000 000.*

References

[1] Artem Babenko and Victor Lempitsky. Efficient indexing of billion-scale datasets of deep descriptors. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[2] L. Best-Rowden and A. K. Jain. Longitudinal study of automatic face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(1):148–162, Jan 2018.

[3] Blumstein, Cohen, Roth, and Visher, editors. *Random parameter stochastic models of criminal careers*. National Academy of Sciences Press, 1986.

[4] Thomas P. Bonczar and Lauren E. Glaze. Probation and parole in the united statesm 2007, statistical tables. Technical report, Bureau of Justice Statistics, December 2008.

[5] White D., Kemp R. I., Jenkins R., Matheson M, and Burton A. M. Passport officers errors in face matching. *PLoS ONE*, 9(8), 2014. e103510. doi:10.1371/journal. pone.0103510.

[6] P. Grother, G. W. Quinn, and P. J. Phillips. Evaluation of 2d still-image face recognition algorithms. NIST Interagency Report 7709, National Institute of Standards and Technology, 8 2010. http://face.nist.gov/mbe as MBE2010 FRVT2010.

[7] P. J. Grother, R. J. Micheals, and P. J. Phillips. Performance metrics for the frvt 2002 evaluation. In *Proceedings of Audio and Video Based Person Authentication Conference (AVBPA)*, June 2003.

[8] Patrick Grother and Mei Ngan. Interagency report 8009, performance of face identification algorithms. *Face Recognition Vendor Test (FRVT)*, May 2014.

[9] Patrick Grother, George Quinn, and Mei Ngan. Face in video evaluation (five) face recognition of non-cooperative subjects. Interagency Report 8173, National Institute of Standards and Technology, March 2017. https://doi.org/10.6028/NIST.IR.8173.

[10] Patrick Grother, George W. Quinn, and Mei Ngan. Face recognition vendor test - still face image and video concept, evaluation plan and api. Technical report, National Institute of Standards and Technology, 7 2013. http://biometrics.nist.gov/cs_links/face/frvt/frvt2012/NIST_FRVT2012_api_Aug15.pdf.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[12] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[13] Masato Ishii, Hitoshi Imaoka, and Atsushi Sato. Fast k-nearest neighbor search for face identification using bounds of residual score. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 194–199, Los Alamitos, CA, USA, May 2017. IEEE Computer Society.

[14] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *CoRR*, abs/1702.08734, 2017.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

[15] Ira Kemelmacher-Shlizerman, Steven M. Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. *CoRR*, abs/1512.00596, 2015.

[16] Yury A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *CoRR*, abs/1603.09320, 2016.

[17] Joyce A. Martin, Brady E. Hamilton, Michelle J.K. Osterman, Anne K. Driscoll, , and Patrick Drake. National vital statistics reports. Technical Report 8, Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System, Division of Vital Statistics, November 2018.

[18] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[19] P. Jonathon Phillips, Amy N. Yates, Ying Hu, Carina A. Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G. Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, Jun-Cheng Chen, Carlos D. Castillo, Rama Chellappa, David White, and Alice J. O'Toole. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018.

[20] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015.

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[22] Jeroen Smits and Christiaan Monden. Twinning across the developing world. *PLOS ONE*, 6(9):1–5, 09 2011.

[23] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[24] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 1701–1708, Washington, DC, USA, 2014. IEEE Computer Society.

[25] A. Towler, R. I. Kemp, and D White. *Unfamiliar face matching systems in applied settings*. Nova Science, 2017.

[26] Working Group 3. Ed. M. Werner. *ISO/IEC 19794-5 Information Technology - Biometric Data Interchange Formats - Part 5: Face image data*. JTC1 :: SC37, 2 edition, 2011. http://webstore.ansi.org.

[27] David White, James D. Dunn, Alexandra C. Schmid, and Richard I. Kemp. Error rates in users of automatic face recognition software. *PLoS ONE*, 10:1–14, October 2015.

[28] Bradford Wing and R. Michael McCabe. Special publication 500-271: American national standard for information systems data format for the interchange of fingerprint, facial, and other biometric information part 1. Technical report, NIST, September 2015. ANSI/NIST ITL 1-2015.

[29] Andreas Wolf. Portrait quality - (reference facial images for mrtd). Technical report, ICAO, April 2018.

[30] D. Yadav, N. Kohli, P. Pandey, R. Singh, M. Vatsa, and A. Noore. Effect of illicit drug abuse on face recognition. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–7, Los Alamitos, CA, USA, mar 2016. IEEE Computer Society.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8271

NISTIR 8280

Face Recognition Vendor Test (FRVT)
Part 3: Demographic Effects

Patrick Grother
Mei Ngan
Kayee Hanaoka

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8280

National Institute of
Standards and Technology
U.S. Department of Commerce

NISTIR 8280

Face Recognition Vendor Test (FRVT)
Part 3: Demographic Effects

Patrick Grother
Mei Ngan
Kayee Hanaoka
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8280

December 2019

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Walter Copan, NIST Director and Undersecretary of Commerce for Standards and Technology

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

National Institute of Standards and Technology Interagency or Internal Report 8280
Natl. Inst. Stand. Technol. Interag. Intern. Rep. 8280, 81 pages (December 2019)

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8280

EXECUTIVE SUMMARY

OVERVIEW This is the third in a series of reports on ongoing face recognition vendor tests (FRVT) executed by the National Institute of Standards and Technology (NIST). The first two reports cover, respectively, the performance of one-to-one face recognition algorithms used for verification of asserted identities, and performance of one-to-many face recognition algorithms used for identification of individuals in photo data bases. This document extends those evaluations to document accuracy variations across demographic groups.

MOTIVATION The recent expansion in the availability, capability, and use of face recognition has been accompanied by assertions that demographic dependencies could lead to accuracy variations and potential bias. A report from Georgetown University [14] work noted that prior studies [22], articulated sources of bias, described the potential impacts particularly in a policing context, and discussed policy and regulatory implications. Additionally, this work is motivated by studies of demographic effects in more recent face recognition [9, 16, 23] and gender estimation algorithms [5, 36].

AIMS AND SCOPE NIST has conducted tests to quantify demographic differences in contemporary face recognition algorithms. This report provides details about the recognition process, notes where demographic effects could occur, details specific performance metrics and analyses, gives empirical results, and recommends research into the mitigation of performance deficiencies.

NIST intends this report to inform discussion and decisions about the accuracy, utility, and limitations of face recognition technologies. Its intended audience includes policy makers, face recognition algorithm developers, systems integrators, and managers of face recognition systems concerned with mitigation of risks implied by demographic differentials.

WHAT WE DID The NIST Information Technology Laboratory (ITL) quantified the accuracy of face recognition algorithms for demographic groups defined by sex, age, and race or country of birth.

We used both one-to-one verification algorithms and one-to-many identification search algorithms. These were submitted to the FRVT by corporate research and development laboratories and a few universities. As prototypes, these algorithms were not necessarily available as mature integrable products. Their performance is detailed in FRVT reports [16, 17].

We used these algorithms with four large datasets of photographs collected in U.S. governmental applications that are currently in operation:

- ▷ **Domestic mugshots** collected in the United States.
- ▷ **Application photographs** from a global population of applicants for immigration benefits.
- ▷ **Visa photographs** submitted in support of visa applicants.
- ▷ **Border crossing photographs** of travelers entering the United States.

All four datasets were collected for authorized travel, immigration or law enforcement processes. The first three sets have good compliance with image capture standards. The last set does not, given constraints on capture duration and environment. Together these datasets allowed us to process a total of 18.27 million images of 8.49 million people through 189 mostly commercial algorithms from 99 developers.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

The datasets were accompanied by sex and age metadata for the photographed individuals. The mugshots have metadata for race, but the other sets only have country-of-birth information. We restrict the analysis to 24 countries in 7 distinct global regions that have seen lower levels of long-distance immigration. While country-of-birth information may be a reasonable proxy for race in these countries, it stands as a meaningful factor in its own right particularly for travel-related applications of face recognition.

The tests aimed to determine whether, and to what degree, face recognition algorithms differed when they processed photographs of individuals from various demographics. We assessed accuracy by demographic group and report on false negative and false positive effects. False negatives are the failure to associate one person in two images; they occur when the similarity between two photos is low, reflecting either some change in the person's appearance or in the image properties. False positives are the erroneous association of samples of two persons; they occur when the digitized faces of two people are similar.

In background material that follows we give examples of how algorithms are used, and we elaborate on the consequences of errors noting that the impacts of demographic differentials can be advantageous or disadvantageous depending on the application.

WHAT WE FOUND

The accuracy of algorithms used in this report has been documented in recent FRVT evaluation reports [16, 17]. These show a wide range in accuracy across developers, with the most accurate algorithms producing many fewer errors. These algorithms can therefore be expected to have smaller demographic differentials.

Contemporary face recognition algorithms exhibit demographic differentials of various magnitudes. Our main result is that false positive differentials are much larger than those related to false negatives and exist broadly, across many, but not all, algorithms tested. Across demographics, false positives rates often vary by factors of 10 to beyond 100 times. False negatives tend to be more algorithm-specific, and vary often by factors below 3.

▷ **False positives:** Using the higher quality Application photos, false positive rates are highest in West and East African and East Asian people, and lowest in Eastern European individuals. This effect is generally large, with a factor of 100 more false positives between countries. However, with a number of algorithms developed in China this effect is reversed, with low false positive rates on East Asian faces. With domestic law enforcement images, the highest false positives are in American Indians, with elevated rates in African American and Asian populations; the relative ordering depends on sex and varies with algorithm.

We found false positives to be higher in women than men, and this is consistent across algorithms and datasets. This effect is smaller than that due to race.

We found elevated false positives in the elderly and in children; the effects were larger in the oldest and youngest, and smallest in middle-aged adults.

- **False negatives:** With domestic mugshots, false negatives are higher in Asian and American Indian individuals, with error rates above those in white and African American faces (which yield the lowest false negative rates). However, with lower-quality border crossing images, false negatives are generally higher in people born in Africa and the Caribbean, the effect being stronger in older individuals. These differing results relate to image quality: The mugshots were collected with a photographic setup specifically standardized to produce high-quality images across races; the border crossing images deviate from face image quality standards.

  In cooperative access control applications, false negatives can be remedied by users making second attempts.

The presence of an enrollment database affords one-to-many identification algorithms a resource for mitigation of demographic effects that purely one-to-one verification systems do not have. Nevertheless, demographic differentials present in one-to-one verification algorithms are usually, but not always, present in one-to-many search algorithms. One important exception is that some developers supplied highly accurate identification algorithms for which false positive differentials are undetectable.

More detailed results are introduced in the Technical Summary.

IMPLICATIONS OF THESE TESTS

Operational implementations usually employ a single face recognition algorithm. Given algorithm-specific variation, it is incumbent upon the system owner to know their algorithm. While publicly available test data from NIST and elsewhere can inform owners, it will usually be informative to specifically measure accuracy of the operational algorithm on the operational image data, perhaps employing a biometrics testing laboratory to assist.

Since different algorithms perform better or worse in processing images of individuals in various demographics, policy makers, face recognition system developers, and end users should be aware of these differences and use them to make decisions and to improve future performance. We supplement this report with more than 1200 pages of charts contained in seventeen annexes that include exhaustive reporting of results for each algorithm. These are intended to show the breadth of the effects, and to inform the algorithm developers.

There are a variety of techniques that might mitigate performance limitations of face recognition systems performance issues overall and specifically those that relate to demographics. This report includes recommendations for research in developing and evaluating the value, costs, and benefits of potential mitigation techniques - see sections 8 and 9.

Reporting of demographic effects often has been incomplete in academic papers and in media coverage. In particular, accuracy is discussed without stating the quantity of interest be it false negatives, false positives or failure to enroll. As most systems are configured with a fixed threshold, it is necessary to report both false negative and false positive rates for each demographic group at that threshold. This is rarely done - most reports are concerned only with false negatives. We make suggestions for augmenting reporting with respect to demographic difference and effects.

BACKGROUND: ALGORITHMS, ERRORS, IMPACTS

FACE ANALYSIS: CLASSIFICATION, ESTIMATION, RECOGNITION

Before presenting results in the Technical Summary we describe what face recognition is, contrasting it with other applications that analyze faces, and then detail the errors that are possible in face verification and identification and their impacts.

Much of the discussion of face recognition bias in recent years cites two studies [5, 36] showing poor accuracy of face gender classification algorithms on black women. Those studies did not evaluate face recognition algorithms, yet the results have been widely cited to indict their accuracy. Our work was undertaken to quantify analogous effects in face recognition algorithms. We strongly recommend that reporting of bias should include information about the class of algorithm evaluated. We use the term **face analysis** as an umbrella for any algorithm that consumes face images and produces some output. Within that are **estimation** algorithms that output some continuous quantity (e.g., age or degree of fatigue). There are **classification** algorithms that aim to determine some categorical quantity such as the sex of a person or their emotional state. Face classification algorithms are built with inherent knowledge of the classes they aim to produce (e.g., happy, sad). Face **recognition** algorithms, however, have no built-in notion of a particular person. They are not built to identify particular people; instead they include a face detector followed by a feature extraction algorithm that converts one or more images of a person into a vector of values that relate to the identity of the person. The extractor typically consists of a neural network that has been trained on ID-labeled images available to the developer. In operations, they act as generic extractors of identity-related information from photos of persons they have usually never seen before. Recognition proceeds as a differential operator: Algorithms compare two feature vectors and emit a similarity score. This is a vendor-defined numeric value expressing how similar the parent faces are. It is compared to a threshold value to decide whether *two* samples are from, or represent, the same person or not. Thus, recognition is mediated by persistent identity information stored in a feature vector (or "template"). Classification and estimation, on the other hand, are single-shot operations from *one* sample alone, employing machinery that is different from that used for face recognition.

VERIFICATION

Errors: A comparison of images from the same person yields a genuine or "mate" score. A comparison of images from different people yields an imposter or "nonmate" score. Ideally, nonmate scores should be low and mate scores should be high. In practice, some imposter scores are above a numeric threshold giving false positives, and some genuine comparisons yield scores below threshold giving false negatives.

Applications: One-to-one verification is used in applications including logical access to a phone or physical access through a security check point. It also supports non-repudiation e.g. to authorize the dispensing of a prescription drug. Two photos are involved: one in the database that is compared with one taken of the person seeking access to answer the question: "Is this the same person or not?"

Impact of errors: Errors have different implications for the system owner and for the individual whose photograph is being used, depending upon the application. In verification applications, false negatives cause inconvenience for the user. For example, an individual may not be able to get into their phone or they are delayed entering a facility or crossing a border. These errors can usually be remediated with a second attempt. False positives, on the other hand, present a security concern to the system owner, as they allow access to imposters.

IDENT-IFICATION

Identification algorithms, referred to commonly as one-to-many or "1-to-N" search algorithms, notionally compare features extracted from a search "probe" image with all feature vectors previously enrolled from "gallery" images. The algorithms return either a fixed number of the most similar candidates, or only those that are above a preset threshold. A candidate is an index and a similarity score. Some algorithms execute an exhaustive search of all N enrollments and a sort operation to yield the most similar. Other algorithms implement "fast-search" techniques [2,19,21,26] that avoid many of the N comparisons and are therefore highly economical [17].

Identification applications: There are two broad uses of identification algorithms. First, they can be used to facilitate positive access like in one-to-one verification but without presentation of an identity claim. For example, a subject is given access to a building solely on the basis of presentation a photograph that matches *any* enrolled identity with a score above threshold. Second, they can be used for so-called negative identification where the system operator claims implicitly that searched individuals are not enrolled - for example, checking databases of gamblers previously banned from a casino.

Impacts: As with verification, the impact of a demographic differential will depend on the application. In one-to-many searches, false positives primarily occur when a search of a subject who is not present in the database yields a candidate identity for human review. This type of "one to many" search is often employed to check for a person who might be applying for a visa or driver's license under a name different than their own. False positives may also occur when a search of someone who is enrolled produces the wrong identity with, or instead of, the correct identity. Identification algorithms produce such outcomes when the search yields a comparison score above a chosen threshold.

In identification applications such as visa or passport fraud detection, or surveillance, a false positive match to another individual could lead to a false accusation, detention or deportation. Higher false negatives would be an advantage to an enrollee in such a system, as their fraud would go undetected, and a disadvantage to the system owner whose security goals will be undermined.

Investigation: This is a special-case application of identification algorithms where the threshold is set to zero so that all searches will produce a fixed number of candidates. In such cases, the false positive identification rate is 100% because any search of someone not in the database will still yield candidates. Algorithms used in this way are part of a hybrid *machine-human system*: The algorithm offers up candidates for human adjudication, for which labor must be available. In such cases, false positive differentials from the algorithm are immaterial - the machine returns say 50 candidates regardless. What matters then is the human response, and the evidence there is for both poor [10, 42] and varied human capability, even without time constraints [34], and sex and race performance differentials, particularly an interaction between the reviewer's demographics with those of the photographs under review [7]. The interaction of machine and human is beyond the scope of this report, as is human efficacy.

TECHNICAL SUMMARY

This section summarizes the results of the study. This is preceded by an introduction to terminology and discussion of a vital aspect in reporting demographic effects, namely that it is necessary to report both false negative and false positive error rates.

ACCURACY DIFFERENTIALS When similarity scores are computed over a collection of images from demographic A (say elderly Asian men) they may be higher than from demographic B (say young Asian women). We adopt terminology from a Department of Homeland Security Science and Technology Directorate article [20] and define **differential performance** as a "difference in the genuine or imposter [score] distributions". Such differentials are inconsequential unless they prompt a **differential outcome.** An outcome occurs when a score is compared with an operator-defined threshold. A genuine score below threshold yields a false negative outcome, and an imposter score at or above threshold, a false positive outcome. The subject of this report is to quantify differential outcomes between demographics. The term demographic differential is inherited from an ISO technical report [6] now under development.

FIXED THRESHOLD OPERATION A crucial point in reasoning about differentials is that the vast majority of biometric systems are configured with a fixed threshold against which all comparisons are made (i.e., the threshold is not tailored to cameras, environmental conditions or, particularly, demographics). Most academic studies ignore this point (even in demographics e.g., [13]) by reporting false negative rates at fixed false positive rates rather than at fixed thresholds, thereby hiding excursions in false positive rates and misstating false negative rates. This report includes documentation of demographic differentials about typical operating thresholds.

We report false positive and false negative rates separately because the consequences of each type of error are of importance to different communities. For example, in a one-to-one access control, false negatives inconvenience legitimate users; false positives undermine a system owners security goals. On the other hand, in a one-to-many deportee detection application, a false negative would present a security problem, and a false positive would flag legitimate visitors. The prior probability of imposters in each case is important. For example, in some access control cases, imposters almost never attempt access and the only germane error rate is the false negative rate.

RESULTS OVERVIEW We found empirical evidence for the existence of demographic differentials in the majority of contemporary face recognition algorithms that we evaluated. The false positive differentials are much larger than those related to false negatives. False positive rates often vary by one or two orders of magnitude (i.e., 10x, 100x). False negative effects vary by factors usually much less than 3. The false positive differentials exist broadly, across many, but not all, algorithms. The false negatives tend to be more algorithm-specific. Research toward mitigation of differentials is discussed in sections 9 and 8.

The accuracy of algorithms used in this report has been documented in recent FRVT evaluation reports [16, 17]. These show a wide range in accuracy across algorithm developers, with the most accurate algorithms producing many fewer errors than lower-performing variants. More accurate algorithms produce fewer errors, and will be expected therefore to have smaller demographic differentials.

FALSE NEGATIVES

With regard to false negative demographic differentials we make the observations below. Note that in real-time cooperative applications, false negatives can often be remedied by making second attempts.

- ▷ False negative error rates vary strongly by algorithm, from below 0.5% to above 10%. For the more accurate algorithms, false negative rates are usually low with average demographic differentials being, necessarily, smaller still. This is an important result: use of inaccurate algorithms will increase the magnitude of false negative differentials. See Figure 22 and Annex 12 .
- ▷ In domestic mugshots, false negatives are higher in Asian and American Indian individuals, with error rates above those in white and black faces. The lowest false negative rates occur in black faces. This result might not be related to race - it could arise due to differences in the time elapsed between photographs because ageing is highly influential on face recognition false negatives. We will report on that analysis going forward. See Figure 17.
- ▷ False negative error rates are often higher in women and in younger individuals, particularly in the mugshot images. There are many exceptions to this, so universal statements pertaining to algorithms false negative rates across sex and age are not supported.
- ▷ When comparing high-quality application photos, error rates are very low and measurement of false negative differentials across demographics is difficult. This implies that better image quality reduces false negative rates and differentials. See Figure 22.
- ▷ When comparing high-quality application images with lower-quality border crossing images, false negative rates are higher than when comparing the application photos. False negative rates are often higher in recognition of women, but the differentials are smaller and not consistent. See Figure 21.
- ▷ In the border crossing images, false negatives are generally higher in individuals born in Africa and the Caribbean, the effect being stronger in older individuals. See Figure 18.

FALSE POSITIVES

Verification Algorithms: With regard to false positive demographic differentials we make the following observations.

- ▷ We found false positives to be between 2 and 5 times higher in women than men, the multiple varying with algorithm, country of origin and age. This increase is present for most algorithms and datasets. See Figure 6.
- ▷ With respect to race, false positive rates are highest in West and East African and East Asian people (but with exceptions noted next). False positive rates are also elevated but slightly less so in South Asian and Central American people. The lowest false positive rates generally occur with East European individuals. See Figure 5.
- ▷ A number of algorithms developed in China give low false positive rates on East Asian faces, and sometimes these are lower than those with Caucasian faces. This observation - that the location of the developer as a proxy for the race demographics of the data they used in training - matters was noted in 2011 [33], and is potentially important to the reduction of demographic differentials due to race and national origin.

- ▷ We found elevated false positives in the elderly and in children; the effects were larger in the oldest adults and youngest children, and smallest in middle aged adults. The effects are consistent across country-of-birth, datasets and algorithms but vary in magnitude. See Figure 14 and Figure 15.
- ▷ With mugshot images, the highest false positives are in American Indians, with elevated rates in African American and Asian populations; the relative ordering depends on sex and varies with algorithm. See Figure 12 and Figure 13.

Identification Algorithms: The presence of an enrollment database affords one-to-many algorithms a resource for mitigation of demographic effects that purely one-to-one verification systems do not have. We note that demographic differentials present in one-to-one verification algorithms are usually, but not always, present in one-to-many search algorithms. See Section 7.

One important exception is that some developers supplied identification algorithms for which false positive differentials are undetectable. Among those is Idemia, who publicly described how this was achieved [15]. A further algorithm, NEC-3, is on many measures, the most accurate we have evaluated. Other developers producing algorithms with stable false positive rates are Aware, Toshiba, Tevian and Real Networks. These algorithms also give false positive identification rates that are approximately independent of the size of enrollment database. See Figure 27.

PRIOR WORK

This report is the first to describe demographic differentials for identification algorithms. There are, however, recent prior tests of verification algorithms whose results comport with ours regarding demographic differentials between races.

- ▷ Using four verification algorithms applied to domestic mugshots, the Florida Institute of Technology and its collaborators showed [23] simultaneously elevated false positives and reduced false negatives in African Americans vs. Caucasians.
- ▷ Cavazos et al. [8] applied four verification algorithms to GBU challenge images [32] to show order-of-magnitude higher false positives in Asians vs. Caucasians. The paper articulates five lessons related to measurement of demographic effects.
- ▷ In addition, a recent Department of Homeland Security (DHS) Science and Technology / SAIC study [20] using a leading commercial algorithm showed that pairing of imposters by age, sex and race gives false positive rates that are two orders of magnitude higher than by pairing individuals randomly.
- ▷ On an approximately monthly schedule starting in 2017, NIST has reported [16] on demographic effects in one-to-one verification algorithms submitted to the FRVT process. Those tests employed smaller sets of mugshot and visa photographs than are used here.

WHAT WE DID NOT DO This report establishes context, gives results and impacts, and discusses additional research that can support mitigation of observed deficiencies. It does not address the following:

▷ **Training of algorithms:** We did not train algorithms. The prototype algorithms submitted to NIST are fixed and were not refined or adapted. This reflects the usual operational situation in which face recognition systems are not adapted on customers local data. We did not attempt, or invite developers to attempt, mitigation of demographic differentials by retraining the algorithms on image sets maintained at NIST. We simply ran the tests using algorithms as submitted.

▷ **Analyze cause and effect:** We did not make efforts to explain the technical reasons for the observed results, nor to build an inferential model of them. Specifically, we have not tried to relate recognition errors to skin tone or any other phenotypes evident in faces in our image sets. We think it likely that modeling will need richer sets of covariates than are available. In particular, efforts to estimate skin tone and other phenotypes will involve an algorithm that itself may exhibit demographic differentials.

We did not yet pursue regression approaches due to the volume of data, the number of algorithms tested, and the need to model each recognition algorithms separately, as they are built and trained independently. Due to their ability to handle imbalanced data, we note, however, the utiltity of mixed effects models [3, 4, 9] previously developed for explaining recognition failure. Such approaches can use subject-specific variables (age, sex, race, etc.) and image-specific variables (contrast, brightness, blur, uniformity, etc.). Models are often useful, even though it is inevitable that germane quantities will be unavailable to the analysis.

▷ **Consider the effect of cameras:** The possible role of the camera, and the subject-camera interaction, has been detailed recently [9]. This is particularly important when standards-compliant photography is not possible, or not intended, for example, in high throughput access control. Without access to human-camera interaction data, we do not report on quantities like satisfaction, difficulty of use, and failure to enroll. Along these lines, it has been suggested [41] that NISTs tests using standards-compliant images "don't translate to everyday scenarios".

In fact, we note demographic effects *even* in high-quality images, notably elevated false positives. Additionally, we quantify false negatives on a border crossing dataset which is collected at a different point in the trade space between quality and speed than are our other three mostly high-quality portrait datasets.

Finally, some governmental organizations dedicated resources to advancing standards so that the "real-world" images in their applications are high-quality portraits. For example, the main criminal justice application is supported by the FBI and others being proactive in the 1990s in establishing portrait capture standards, and then promulgating them.

▷ **Use wild images:** We did not use image data from the Internet nor from video surveillance. This report does not capture demographic differentials that may occur in such photographs.

RESEARCH RECOMMENDATIONS

We now discuss research germane to the quantification, handling and mitigation of demographic differentials.

Testing: Since 2017 NIST has provided demographic differential data to developers of one-to-one verification algorithms. Our goal has been to encourage developers to remediate the effects. While that may have happened in some cases, a prime incentive for a developer when participating in NIST evaluations is to reduce false negatives rates globally. Going forward, we plan to start reporting accuracy that pushes developers to produce approximately equal false positive rates across all demographics.

Mitigation of false positive differentials: With adequate research and development, the following may prove effective at mitigating demographic differentials with respect to false positives: Threshold elevation, refined training, more diverse training data, discovery of features with greater discriminative power - particularly techniques capable of distinguishing between twins - and use of face and iris as a combined modality. These are discussed in section 9. We also discuss, and discount, the idea of user-specific thresholds.

Mitigation of false negative differentials: False negative error rates, and demographic differentials therein, are reduced in standards-compliant images. This motivates the suggestions of further research into image quality analysis, face-aware cameras and improved standards-compliance discussed in section 8.

Policy research: The degree to which demographic differentials could be tolerated has never been formally specified in any biometric application. Any standard directed toward limiting allowable differentials in the automated processing of digitized biological characteristics might weigh the actual consequences of differentials which are strongly application dependent.

REPORTING OF DEMOGRAPHIC EFFECTS

Reporting of demographic effects has been incomplete, in both academic papers and in media coverage. In particular, accuracy is discussed without specifying, particularly, false positives or false negatives. We therefore suggest that reports covering demographic differentials should describe:

- ▷ The purpose of the system - initial enrollment of individuals into a system, identity verification or identification;
- ▷ The stage at which the differential occurred - at the camera, during quality assessment, in the detection and feature extraction phase, or during recognition;
- ▷ The relevant metric: false positive or false negative occurrences during recognition, failures to enroll, failed detections by the camera, for example;
- ▷ Any differentials in duration of processes or difficulty in using the system;
- ▷ Any known information on recognition threshold value, whether the threshold is fixed, and what the target false positive rate is;
- ▷ Which demographic group has the elevated failure rates - for example by age, sex, race, height, or in some intersection thereof; and
- ▷ Consequences of any error, if known, and procedures for error remediation.

ACKNOWLEDGMENTS The authors are grateful to Yevgeniy Sirotin and John Howard of SAIC at the Maryland Test Facility, Arun Vemury of DHS S&T, Michael King of Florida Institute of Technology, and John Campbell of Bion Biometrics for detailed discussions of their work in this area.

The authors are grateful to staff in the NIST Biometrics Research Laboratory for infrastructure supporting rapid evaluation of algorithms.

DISCLAIMER Specific hardware and software products identified in this report were used in order to perform the evaluations described in this document. In no case does identification of any commercial product, trade name, or vendor, imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose. Developers participating in FRVT grant NIST permission to publish evaluation results.

ANNEXES

We supplement this report with more than 1200 pages of charts contained in 17 Annexes which include exhaustive reporting of results for each algorithm. These are intended to show the breadth of the effects and to inform the algorithms' developers. We do not take averages over algorithms, for example the average increase of false match rate in women, because averages of samples from different distributions are seldom meaningful (by analogy, taking the average of temperatures in Montreal and Miami). Applications typically employ just one algorithm, so averages and indeed any statements purporting to summarize the entirety of face recognition will not always be correct.

The annexes to this report are listed in Table 1. The first four detail the datasets used in this report. The remaining annexes contain more than 1200 pages of automatically generated graphs, usually one for each algorithm evaluated. These are intended to show the breadth of the effects, and to inform the algorithms' developers.

#	CATEGORY		DATASET	CONTENT
Annex 1	Datasets		Mugshot	Description and examples of images and metadata: Mugshots
Annex 2	Datasets		Application	Description and examples of images and metadata: Application portraits
Annex 3	Datasets		Visa	Description and examples of images and metadata: Visa portraits
Annex 4	Datasets		Border crossing	Description and examples of images and metadata: Border crossing photos
Annex 5	Results	1:1	Application	False match rates for demographically matched impostors
Annex 6	Results	1:1	Mugshot	Cross-race and sex false match rates in United States mugshot images
Annex 7	Results	1:1	Application	Cross-race and sex false match rates in worldwide application images
Annex 8	Results	1:1	Application	False match rates with matched demographics using application images
Annex 9	Results	1:1	Application	Cross-age false match rates with application photos
Annex 10	Results	1:1	Visa	Cross age false match rates with visa photos
Annex 11	Results	1:1	Mugshot	Cross age and country with application photos
Annex 12	Results	1:1	Mugshot	Error tradeoff characteristics with United States mugshots
Annex 13	Results	1:1	Mugshot	False negative rates in United States mugshot images by sex and race
Annex 14	Results	1:1	Mugshot	False negative rates by country for global application and border crossing photos
Annex 15	Results	1:1	Mugshot	Genuine and impostor score distributions for United States mugshots
Annex 16	Results	1:N	Mugshot	Identification error characteristics by race and sex
Annex 17	Results	1:N	Mugshot	Candidate list score magnitudes by sex and race

Table 1: **Annexes and their content.**

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

TERMS AND DEFINITIONS

The following table defines common terms appearing in this document. A more complete, consistent biometrics vocabulary is available as ISO/IEC 2382 Part 37.

DATA TYPES	Feature vector	A vector of real numbers that encodes the identity of a person
	Sample	One or more images of the face of a person
	Similarity score	Degree of similarity of two faces in two samples, as rendered by a recognition algorithm
	Template	Data produced by face recognition algorithm that includes a feature vector
	Threshold	Any real number, against which similarity scores are compared to produce a verification decision
ALGORITHM COMPONENTS	Face detector	Component that finds faces in an image
	Comparator	Component that compares two templates and produces a similarity score
	Searcher	Component that searches a database of templates to produce a list of candidates
	Template generator	Component of a face recognition algorithm that converts a sample into a template; this component implicitly embeds a face detector
ONE-TO-ONE VERIFICATION	Imposter comparison	Comparison of samples from different persons
	Genuine comparison	Comparison of samples from the same person
	False match	Incorrect association of two samples from different persons, declared because similarity score is at or above a threshold
	False match rate	Proportion of imposter comparisons producing false matches
	False non-match	Failure to associate two samples from one person, declared because similarity score is below a threshold
	False non-match rate	Proportion of genuine comparisons producing false non-matches
	Verification	The process of comparing two samples to determine if they belong to the same person or not
ONE-TO-MANY IDENTIFICATION	Gallery	A set of templates, each tagged with an identity label
	Consolidated gallery	A gallery for which all samples of a person are enrolled under one identifier, whence $N = N_G$
	Unconsolidated gallery	A gallery for which samples of a person are enrolled under different identifiers, when $N < N_G$
	Identity label	Some index or pointer to an identifier for an individual
	Identification	The process of searching a probe into gallery
	Identification decision	The assignment either of an identity label or a declaration that a person is not in the gallery
SYMBOLS	FMR	Verification false match rate (measured over comparison of samples)
	FNMR	Verification false non-match rate (measured over comparison of samples)
	FPIR	Identification false match rate (measured over comparison of samples)
	FNIR	Identification false non-match rate (measured over comparison of samples)
	N	The number of subjects whose faces are enrolled into a gallery
	N_G	The number of samples enrolled into a gallery, $N_G \geq N$.
	N_{NM}	The number of non-mated searches conducted
	N_M	The number of mated searches conducted

Contents

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

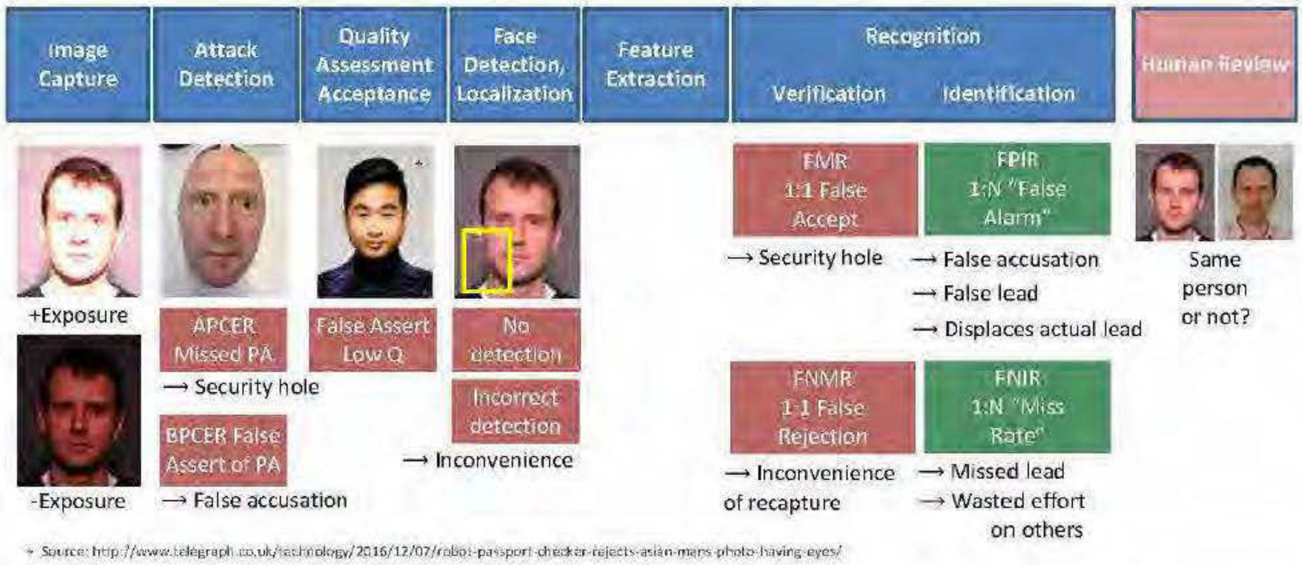


Figure 1: The figure is intended to show possible stages in a face recognition pipeline at which demographic differentials could, in principle, arise. Note that none of these stages necessarily includes algorithms that may be labelled artificial intelligence, though typically the detection and feature extraction modules are AI-based now.

1 Introduction

Over the last two years there has been expanded coverage of face recognition in the popular press. In some part this is due to the expanded capability of the algorithms, a larger number of applications, lowered barriers to algorithm development[1], and, not least, reports that the technology is somehow biased. This latter aspect is based on Georgetown [14] and two reports by MIT [5, 36]. The Georgetown work noted prior studies [22] articulated sources of bias, and described the potential impacts particularly in a policing context, and discussed policy and regulatory implications. The MIT work did not study face recognition, instead it looked at how well publicly accessible cloud-based *estimation* algorithms can determine gender from a single image. The studies have widely cited as evidence that face *recognition* is biased.

This stems from a confusion in terminology: Face classification algorithms, of the kind MIT reported on, accept one face image sample and produce an estimate of age, or sex, or some other property of the subject. Face recognition algorithms, on the other hand, operate as differential operators: They compare identity information in features vectors extract from two face image samples and produce a measure of similarity between the two, which can be used to answer the "question same person or not?". Face algorithms, both one-to-one identity verification and one-to-many search algorithms, are built on this differential comparison. The salient point, in the demographic context, is that one or two people are involved in a comparison and, as we will see, the age,

[1] Gains in face recognition performance stem from well-capitalized AI research in industry and academic leading to the development of convolutional neural networks, and open-source implementations thereof (Caffe, Tensorflow etc.). For face recognition the availability of large numbers of identity-labeled images (from the web, and in the form of web-curated datasets [VGG2, IJB-C]), and the availability of ever more powerful GPUs has supported training those networks.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

sex, race and other demographic properties of both will be material to the recognition outcome.

The MIT reports nevertheless serve as a cautionary tale in two respects. First, that demographic group membership can have a sizeable effect on algorithms that process face photographs; second, that algorithm capability varies considerably by developer.

1.1 Potential sources of bias in face recognition systems

Lost in the discussion of bias is specificity on exactly what component of the process is at fault. Accordingly, we introduce Figure 1 to show that a face recognition system is composed of several parts. The figure shows a notional face recognition pipeline consisting of a capture subsystem, primarily a camera, followed by a presentation attack detection (PAD) module intended to detect impersonation attempts, a quality acceptance (QA) step aimed at checking portrait standard compliance, then the recognition components of feature extraction and 1:1 or 1:N comparison, the output of which may prompt human involvement. The order of the components may be different in some systems, for example the QA component may be coupled to the capture process and would precede PAD. Some components may not exist in some systems, particularly the QA and PAD functions may not be necessary.

The Figure shows performance metrics, any of which could notionally have a demographic differential. Errors at one stage will generally have downstream consequences. In a system where subjects make cooperative presentation to the camera, a person could be rejected in the early stages before recognition itself. For example, a camera equipped with optics that have too narrow a field of view could produce an image of a tall individual in which in which the top part of the head was cropped. This could cause rejection at almost any stage and a system owner would need to determine the origin of errors.

1.2 The role of image quality

Recent research [9] has shown that cameras can have an effect on a generic downstream recognition engine. A poor image can undermine detection or recognition, and it is possible that certain demographics yield photographs ill-suited to face recognition e.g. young children [28], or very tall individuals. As pointed out above there is potential for demographic differentials to appear at the capture stage, that is when only a single image is being collected before any comparison with other images. Demographic differentials that occur during collection could arise from (at least) inadequacies of the camera, from the environment or "stage", and from client-side detection or quality assessment algorithms. Note that manifestly poor (and unrecognizable) images can be collected from mis-configured cameras, without any algorithmic or AI culpability. Indeed, after publication of the MIT studies [5, 36] on bias in gender-estimation algorithms, suspicion fell upon the presence of poor photographs, due to under-exposure of dark-skinned individuals in that dataset. An IBM gender estimation algorithm had been faulted in the MIT study; in response, and previously, IBM has been active in addressing AI bias. Relevant here is that it produced a better algorithm[2], and examined whether skin tone itself drove gender classification accuracy [30, 31] - in short, "skin type *by itself* has a minimal effect on the classification decision".

False negatives occur in biometric systems when samples from one individual yield a comparison score below a threshold. This will occur when the features extracted from two input photographs are insufficiently similar. Recall that face recognition is implemented as a differential operator: two samples are analyzed and compared. So a false negative occurs when two from the same face appear different to the algorithm.

It is very common to attribute false negatives to factors such as pose, illumination and expression so much so that dedicated databases have been built up to support development of algorithms with immunity to such[3]. Invariance to such "nuisance" factors has been the focus of the bulk of face recognition research for more two decades. Indeed over the last five years there have been great advances in this respect due to the adoption of deep convolutional neural networks which demonstrate remarkable tolerance to very sub-standard photographs i.e. those that deviate from formal portrait standards most prominently ISO/IEC 39794-5 and its law-enforcement equivalent ANSI/NIST ITL 1-2017.

However, here we need to distinguish between factors that are expected to affect one photo in a mated pair - due to poor photography (e.g. mis-focus), poor illumination (e.g. too dark), and poor presentation (e.g. head down) - and those that would affect both photographs over time, potentially including properties related to demographics.

1.3 Photographic Standards

In the late 1990s the FBI asked NIST to establish photographic best-practices for mugshot collection[4]. This was done to guide primarily state and local police departments in the capture of photographs that would support forensic (i.e. human) review. It occurred more than a decade before the FBI deployed automated face recognition. That standardization work was conducted in anticipation of digital cameras[5] being available to replace film cameras that had been used for almost a century. The standardization work included consideration of cameras, lights and geometry[6]. There was explicit consideration of the need to capture images of both dark and light skinned individuals, it being understood that it is relatively easy to produce photographs for which

[2] See Mitigating Bias in AI Models.
[3] The famous PIE databases, for example.
[4] Early documents, such as Best Practice Recommendation for the Capture of Mugshots, 1999, seeded later formal standardization of ISO/IEC 19794-5.
[5] See NIST Interagency Report 6322, 1999.
[6] See this overview.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

large areas of dark or bright pixels can render detection of anatomical features impossible.

Face recognition proceeds as a differential operation on features extracted from two photographs. Accuracy can be undermined by poor photography/illumination/presentation and by differences in those i.e. any change in the digital facial appearance. Of course an egregiously underexposed photograph will have insufficient information content, but two photographs taken with even moderately poor exposure can match, and leading contemporary algorithms are highly tolerant of quality degradations.

1.4 Age and ageing

Ageing will change appearance over decades and will ultimately undermine automated face recognition[7]. In the current study, we don't consider ageing to be a demographic factor because it is a slow, more-or-less graceful, process that happens to all of us. However, there is at least one demographic group that ages more quickly than others - children - who are disadvantaged in many automated border control systems either by being excluded by policy, or by encountering higher false negatives. Age itself is a demographic factor as accuracy in the elderly and the young differ for face recognition (usually) and also for fingerprint authentication. This applies even without significant time lapse between two photographs.

Clearly injury or disease can change appearance on very short timescales, so such factors should be excluded, when possible, from studies dedicated to detection of broad demographic effects. Development of equipment and algorithms, and studies thereof, that are dedicated to the inclusive use of biometrics are valuable of course - for example recognition of photosensitive subjects wearing sunglasses, or finger amputees presenting fingerprints.

[7]See recent results for verification algorithms in the FRVT reports, and for identification algorithms in NIST Interagency Report 8271 [17]. For a formal longitudinal analysis of ageing, using mixed-effects models, see Best-Rowden [3].

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

#	SOURCE	IMAGE	NUMBER OF SUBJECTS	NUMBER OF IMAGES	DISCUSSION
1	Cavazos et al. [8] at UT Dallas	Notre Dame GBU [32] portraits	389	<1085	The study showed order-of-magnitude elevations in false positive rates in university volunteer Asian vs. Caucasian faces. The study reported FMR(T). As the study showed neither FNMR(T) nor linked error tradeoff characteristics the false negative differential is not apparent. It discusses the effect of "yoking" i.e the pairing of imposters by sex and race. It deprecates area-under-the-curve (AUC). The study used two related algorithms from the University of Maryland, one open-source algorithm [38], and one older inaccurate pre-DCNN algorithm.
2	Krishnapriya et al. [23] at Florida Inst. Tech	Operational mugshots: Morph db [37]	10350 African Am. + 2769 Caucasians	42620 African Am. + 10611 Caucasians	The study reported: order-of-magnitude elevated false positives in African Americans vs. Caucasians; lower false negative rates in African Americans; and reduced differentials in higher quality images [23,24]. That study used three open-source algorithms, and one commercial algorithm. Two of the open-source algorithms are quite inaccurate and not representative of commercial deployment. Importantly, the study also noted the inadequacies of error tradeoff characteristics for documenting fixed-threshold demographic differentials.
3	Howard et al. [20] at SAIC/MdTF with DHS S&T	Lab collected, adult volunteers [9]	363	-	The study establish useful definitions for "differential performance" and "differential outcome" and for broad and narrow heterogeneity of imposter distributions. It showed order-of-magnitude variation in false positive rates with age, sex and race, establishing an information gain approach to formally ordering their effect. The study employed images from 11 capture devices, and applied one leading commercial verification algorithm.

Table 2: Prior studies.

2 Prior work

All prior work relates to one-to-one verification algorithms. This report, in contrast, includes results for many recent, mostly commercial, algorithms implementing both verification and identification.

Except as detailed below, this report is the first to properly report and distinguish between false positive and false negative effects, something that is often missing in other reports.

The broad effects given in this report concerning age and sex have been known as far back as 2003 [35]. Since 2017, our ongoing FRVT report [16] has reported large false positive differential across sex, age and race.

Tables 2 and 3 summarize recent work in demographic effects in automated face recognition.

#	SOURCE	IMAGE	NUMBER OF SUBJECTS	NUMBER OF IMAGES	DISCUSSION
4	Cook et al. [9] at SAIC/MdTF with DHS S&T	Lab collected, adult volunteers	525		The study deployed mixed-effects regression models to examine dependence of genuine similarity scores on sex, age, height, eyewear, skin reflectance and on capture device. The report displayed markedly different images of the same people from different capture devices, showing potential for the camera to induce demographic differential performance. The study found lower similarity scores in those identifying as Black or African American, comporting with [22] but contrary to the best ageing study [3]. The study also showed that comparison of samples collected on the same day have different demographic differentials than those collected up to four years apart, in particular that women give lower genuine scores than men with time separation. Same-day biometrics are useful for short-term recognition applications like transit through an airport.
5	El Khiyari et al. [13]	Operational mugshots: Morph db [37]	724 adult, balanced on race + sex	2896 = 1448 each African Am. + Caucasians, balanced on sex	The paper used a subset of the MORPH database with two algorithms([38], modified and one COTS) to show better verification error rates in the men, the elderly, and in whites. The study should be discounted for two reasons: First the algorithms give high error rates at very modest false match rates: the best FNMR = 0.06 at FMR = 0.01. Second the paper reports FNMR at fixed FMR, not at fixed thresholds thereby burying FMR differentials. Moreover, the paper does not disclose how imposters were paired e.g. randomly or, say, with same age, race, and sex.

Table 3: Prior studies (continued).

3 Performance metrics

Both verification and identification systems generally commit two kinds of errors, the so-called Type I error where an individual is incorrectly associated with another, and Type II where the individual is incorrectly not associated with themselves.

The ISO/IEC 19795-1 performance testing and reporting standard requires different metrics to be reported for identification and verification implementations. Accordingly the following subsections define the formal metrics used throughout this document.

3.1 Verification metrics

Verification accuracy is estimated by forming two sets of scores: Genuine scores are produced from mated pairs; imposter scores are produced from non-mated pairs. These comparisons should be done in random order so that the algorithm under test cannot infer that a comparison is mated or not.

From a vector of N genuine scores, u, the false non-match rate (FNMR) is computed as the proportion below some threshold, T:

$$\mathrm{FNMR}(T) = 1 - \frac{1}{N}\sum_{i=1}^{N} H(u_i - T) \tag{1}$$

where $H(x)$ is the unit step function, and $H(0)$ taken to be 1.

Similarly, given a vector of M imposter scores, v, the false match rate (FMR) is computed as the proportion above T:

$$\mathrm{FMR}(T) = \frac{1}{M}\sum_{i=1}^{M} H(v_i - T) \tag{2}$$

The threshold, T, can take on any value. We typically generate a set of thresholds from quantiles of the observed imposter scores, v, as follows. Given some interesting false match rate range, $[\mathrm{FMR}_L, \mathrm{FMR}_U]$, we form a vector of K thresholds corresponding to FMR measurements evenly spaced on a logarithmic scale. This supports plotting of FMR on a logarithmic axis. This is done because typical operations target false match rates spanning several decades 10^{-6} to as high as 10^{-2}.

$$T_k = Q_v(1 - \mathrm{FMR}_k) \tag{3}$$

where Q_v is the quantile function, and FMR_k comes from

$$\log_{10} \mathrm{FMR}_k = \log_{10} \mathrm{FMR}_L + \frac{k}{K}\left[\log_{10} \mathrm{FMR}_U - \log_{10} \mathrm{FMR}_L\right] \tag{4}$$

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

Error tradeoff characteristics are plots of FNMR(T) vs. FMR(T). These are plotted with $\mathrm{FMR}_U \rightarrow 1$ and FMR_L as low as is sustained by the number of imposter comparisons, M. This should be somewhat higher than the "rule of three" limit $3/N$ because samples are generally not independent due to the use of the same image in multiple comparisons.

3.2 Identification metrics

Identification accuracy is estimated from two sets of candidate lists: First, a set of candidate lists obtained from mated-searches; second, a set from non-mated searches. These searches should not be conducted by randomly ordering mated and non-mated searches so that the algorithm under test cannot infer that a search has a mate or not. Tests of open-set biometric identification algorithms must quantify frequency of two error conditions:

- ▷ **False positives**: Type I errors occur when search data from a person who has never been seen before is incorrectly associated with one or more enrollees' data.
- ▷ **Misses**: Type II errors arise when a search of an enrolled person's biometric does not return the correct identity.

Many practitioners prefer to talk about "hit rates" instead of "miss rates" - the first is simply one minus the other as detailed below. Sections 3.2.1 and 3.2.2 define metrics for the Type I and Type II performance variables.

Additionally, because recognition algorithms sometimes fail to produce a template from an image, or fail to execute a one-to-many search, the occurrence of such events must be recorded. Further because algorithms might elect to not produce a template from, for example, a poor quality image, these failure rates must be combined with the recognition error rates to support algorithm comparison. This is addressed in section 3.4.

3.2.1 Quantifying false positives

It is typical for a search to be conducted into an enrolled population of N identities, and for the algorithm to be configured to return the closest L candidate identities. These candidates are ranked by their score, in descending order, with all scores required to be greater than or equal to zero. A human analyst might examine either all L candidates, or just the top $R \leq L$ identities, or only those with score greater than threshold, T.

From the candidate lists, we compute **false positive identification rate** as the proportion of non-mate searches that erroneously return candidates:

$$\mathrm{FPIR}(N,T) = \frac{\text{Num. non-mate searches with one or more candidates returned with score at or above threshold}}{\text{Num. non-mate searches attempted.}} \tag{5}$$

Under this definition, FPIR can be computed from the highest non-mate candidate produced in a search - it is not necessary to consider candidates at rank 2 and above. An alternative quantity, selectivity, accounts for multiple candidates above threshold - see [17].

3.2.2 Quantifying hits and misses

If L candidates are returned in a search, a shorter candidate list can be prepared by taking the top $R \leq L$ candidates for which the score is above some threshold, $T \geq 0$. This reduction of the candidate list is done because thresholds may be applied, and only short lists might be reviewed (according to policy or labor availability, for example). It is useful then to state accuracy in terms of R and T, so we define a "miss rate" with the general name **false negative identification rate** (FNIR), as follows:

$$\text{FNIR}(N,R,T) = \frac{\text{Num. mate searches with enrolled mate found outside top R ranks or score below threshold}}{\text{Num. mate searches attempted.}} \tag{6}$$

This formulation is simple for evaluation in that it does not distinguish between causes of misses. Thus a mate that is not reported on a candidate list is treated the same as a miss arising from face finding failure, algorithm intolerance of poor quality, or software crashes. Thus if the algorithm fails to produce a candidate list, either because the search failed, or because a search template was not made, the result is regarded as a miss, adding to FNIR.

Hit rates, and true positive identification rates: While FNIR states the "miss rate" as how often the correct candidate is either not above threshold or not at good rank, many communities prefer to talk of "hit rates". This is simply the **true positive identification rate**(TPIR) which is the complement of FNIR giving a positive statement of how often mated searches are successful:

$$\text{TPIR}(N,R,T) = 1 - \text{FNIR}(N,R,T) \tag{7}$$

This report does not report true positive "hit" rates, preferring false negative miss rates for two reasons. First, costs rise linearly with error rates. For example, if we double FNIR in an access control system, then we double user inconvenience and delay. If we express that as decrease of TPIR from, say 98.5% to 97%, then we mentally have to invert the scale to see a doubling in costs. More subtly, readers don't perceive differences in numbers near 100% well, becoming inured to the "high nineties" effect where numbers close to 100 are perceived indifferently.

Reliability is a corresponding term, typically being identical to TPIR, and often cited in automated (fingerprint) identification system (AFIS) evaluations.

An important special case is the **cumulative match characteristic**(CMC) which summarizes accuracy of mated-searches only. It ignores similarity scores by relaxing the threshold requirement, and just reports the fraction of mated searches returning the mate at rank R or better.

$$\mathrm{CMC}(N, R) = 1 - \mathrm{FNIR}(N, R, 0) \tag{8}$$

We primarily cite the complement of this quantity, $\mathrm{FNIR}(N, R, 0)$, the fraction of mates *not* in the top R ranks.

The **rank one hit rate** is the fraction of mated searches yielding the correct candidate at best rank, i.e. CMC(N, 1). While this quantity is the most common summary indicator of an algorithm's efficacy, it is not dependent on similarity scores, so it does not distinguish between strong (high scoring) and weak hits. It also ignores that an adjudicating reviewer is often willing to look at many candidates.

3.3 DET interpretation

In biometrics, a false negative occurs when an algorithm fails to match two samples of one person – a Type II error. Correspondingly, a false positive occurs when samples from two persons are improperly associated – a Type I error.

Matches are declared by a biometric system when the native comparison score from the recognition algorithm meets some threshold. Comparison scores can be either similarity scores, in which case higher values indicate that the samples are more likely to come from the same person, or dissimilarity scores, in which case higher values indicate different people. Similarity scores are traditionally computed by fingerprint and face recognition algorithms, while dissimilarities are used in iris recognition. In some cases, the dissimilarity score is a distance possessing metric properties. In any case, scores can be either mate scores, coming from a comparison of one person's samples, or nonmate scores, coming from comparison of different persons samples.

The words "genuine" or "authentic" are synonyms for mate, and the word "imposters"" is used as a synonym for nonmate. The words "mate" and "nonmate" are traditionally used in identification applications (such as law enforcement search, or background checks) while genuine and imposter are used in verification applications (such as access control).

An error tradeoff characteristic represents the tradeoff between Type II and Type I classification errors. For identification this plots false negative vs. false positive identification rates i.e. FNIR vs. FPIR parametrically with T. Such plots are often called detection error tradeoff (DET) characteristics or receiver operating characteristic (ROC). These serve the same function – to show error tradeoff – but differ, for example, in plotting the complement of an error rate (e.g. TPIR = 1 − FNIR) and in transforming the axes, most commonly using logarithms, to show multiple decades of FPIR.

3.4 Failure to extract features

During enrollment some algorithms fail to convert a face image to a template. The proportion of failures is the failure-to-enroll rate, denoted by FTE. Similarly, some search images are not converted to templates. The corresponding proportion is termed failure-to-extract, denoted by FTX. We do not report FTX because we assume that the same underlying algorithm is used for template generation for enrollment and search.

In verification, we do not need to explicitly include failure to extract rates into the FNMR and FMR accuracy statements, because we regard any comparison that involves an image for which a failure-to-extract occurred as producing a zero similarity score. This increases FNMR and decreases FMR. Gaming opportunities that theoretically arise from this treatment of FMR are generally not of concern because the algorithm under test does not know whether any given image will be used in genuine comparisons, imposter comparisons or both.

For identification, we similarly incorporate failure-to-extract events into FNIR and FPIR measurements as follows.

- ▷ **Enrollment templates:** Any failed enrollment is regarded as producing a zero length template. Algorithms are required by the API [18] to transparently process zero length templates. The effect of template generation failure on search accuracy depends on whether subsequent searches are mated, or non-mated: Mated searches will fail giving elevated FNIR; non-mated searches will not produce false positives so, to first order, FPIR will be reduced by a factor of $1 - \mathrm{FTE}$.
- ▷ **Search templates and 1:N search**: In cases where the algorithm fails to produce a search template from input imagery, the result is taken to be a candidate list whose entries have no hypothesized identities and zero score. The effect of template generation failure on search accuracy depends on whether searches are mated, or non-mated: Mated searches will fail giving elevated FNIR; Non-mated searches will not produce false positives, so FPIR will be reduced.

This approach is the correct treatment for positive-identification applications such as access control where cooperative users are enrolled and make attempts at recognition. This approach is not appropriate to negative identification applications, such as visa fraud detection, in which hostile individuals may attempt to evade detection by submitting poor quality samples. In those cases, template generation failures should be investigated as though a false alarm had occurred.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

	Developer	Verification algorithms	Identification algorithms
1	3Divi	3divi-003 3divi-004	3divi-0 3divi-3
2	Adera Global PTE Ltd	adera-001	
3	Alchera Inc	alchera-000 alchera-001	alchera-0
4	Alivia / Innovation Sys	isystems-001 isystems-002	isystems-0 isystems-3
5	AllGoVision	allgovision-000	allgovision-000
6	AlphaSSTG	alphaface-001	
7	Amplified Group	amplifiedgroup-001	
8	Anke Investments	anke-004	anke-0 anke-002
9	AnyVision	anyvision-002 anyvision-004	
10	Aware	aware-003 aware-004	aware-0 aware-3
11	Awidit Systems	awiros-001	
12	Ayonix	ayonix-000	ayonix-0
13	Beijing Vion Technology Inc	vion-000	
14	Bitmain	bm-001	
15	CSA IntelliCloud Technology	intellicloudai-001	
16	CTBC Bank Co Ltd	ctbcbank-000	
17	Camvi Technologies	camvi-002 camvi-004	camvi-1 camvi-3 camvi-4
18	China Electronics Import-Export Corp	ceiec-001 ceiec-002	
19	China University of Petroleum	upc-001	
20	Chunghwa Telecom Co. Ltd	chtface-001	
21	Cognitec Systems GmbH	cognitec-000 cognitec-001	cognitec-0 cognitec-2
22	Cyberextruder	cyberextruder-001 cyberextruder-002	
23	Cyberlink Corp	cyberlink-002 cyberlink-003	
24	DSK	dsk-000	
25	Dahua Technology Co Ltd	dahua-002 dahua-003	dahua-0 dahua-1 dahua-002
26	Deepglint	deepglint-001	deepglint-001
27	Dermalog	dermalog-005 dermalog-006	dermalog-0 dermalog-5 dermalog-6
28	DiDi ChuXing Technology Co	didiglobalface-001	
29	Digital Barriers	digitalbarriers-002	
30	Eyedea Recognition		eyedea-0 eyedea-3
31	FaceSoft Ltd	facesoft-000	
32	FarBar Inc	f8-001	f8-001
33	Gemalto Cogent	cogent-003 cogent-004	
34	Glory Ltd	glory-001	glory-0
35	Gorilla Technology	gorilla-003	gorilla-0
36	Guangzhou Pixel Solutions Co Ltd	pixelall-002	pixelall-002
37	Hengrui AI Technology Ltd	hr-001 hr-002	
38	Hikvision Research Institute	hik-001	hik-0 hik-5
39	ID3 Technology	id3-003 id3-004	
40	ITMO University	itmo-005 itmo-006	
41	Idemia	idemia-004 idemia-005	idemia-0 idemia-4 idemia-5
42	Imagus Technology Pty Ltd	imagus-000	imagus-0
43	Imperial College London	imperial-000 imperial-002	imperial-000
44	Incode Technologies Inc	incode-004	incode-0 incode-004

Table 4: Algorithms evaluated in this report.

	Developer	Verification algorithms	Identification algorithms
45	Innovatrics	innovatrics-004 innovatrics-006	innovatrics-0
46	Institute of Information Technologies	iit-001	
47	Intel Research Group	intelresearch-000	
48	Intellivision	intellivision-001 intellivision-002	
49	Is It You	isityou-000	
50	Kakao Corp	kakao-001 kakao-002	
51	Kedacom International Pte	kedacom-000	kedacom-001
52	Kneron Inc	kneron-003	
53	Lomonosov Moscow State University	intsysmsu-000	intsysmsu-000
54	Lookman Electroplast Industries	lookman-002 lookman-004	
55	Megvii/Face++	megvii-001 megvii-002	megvii-0 megvii-1
56	MicroFocus	microfocus-002 microfocus-001	microfocus-0
57	Microsoft		microsoft-0 microsoft-5
58	Momentum Digital Co Ltd	sertis-000	
59	Moontime Smart Technology	mt-000	
60	N-Tech Lab	ntechlab-006 ntechlab-007	ntechlab-0 ntechlab-6 ntechlab-007
61	NEC		nec-2 nec-3
62	Neurotechnology	neurotechnology-005 neurotechnology-006	neurotechnology-0 neurotechnology-5 neurotechnology-007
63	Nodeflux	nodeflux-001 nodeflux-002	
64	NotionTag Technologies Private Limited	notiontag-000	
65	Panasonic R+D Center Singapore	psl-002 psl-003	
66	Paravision (EverAI)	everai-paravision-003 paravision-004	everai-0 everai-3 everai-paravision-004
67	Rank One Computing	rankone-007	rankone-0 rankone-5 rankone-006 rankone-007
68	Realnetworks Inc	realnetworks-002 realnetworks-003	realnetworks-0 realnetworks-2 realnetworks-003
69	Remark Holdings	remarkai-001	remarkai-0 remarkai-000
70	Rokid Corporation Ltd	rokid-000	
71	Saffe Ltd	saffe-001 saffe-002	
72	Sensetime Group Ltd	sensetime-002	sensetime-0 sensetime-1 sensetime-002
73	Shaman Software	shaman-000 shaman-001	shaman-0
74	Shanghai Jiao Tong University	sjtu-001	
75	Shanghai Ulucu Electronics Technology Co. Ltd	uluface-002	
76	Shanghai Universiy - Shanghai Film Academy	shu-001	
77	Shanghai Yitu Technology	yitu-003	yitu-0 yitu-4 yitu-5
78	Shenzhen EI Networks Limited	einetworks-000	
79	Shenzhen Inst Adv Integrated Tech CAS	siat-004 siat-002	siat-0
80	Shenzhen Intellifusion Technologies Co Ltd	intellifusion-001	
81	Smilart	smilart-002 smilart-003	smilart-0
82	Star Hybrid Limited	starhybrid-001	
83	Synesis	synesis-005	synesis-0
84	Tech5 SA	tech5-002 tech5-003	tech5-001
85	Tencent Deepsea Lab	deepsea-001	deepsea-001
86	Tevian	tevian-004 tevian-005	tevian-0 tevian-4
87	Thales		cogent-0 cogent-3
88	TigerIT Americas LLC	tiger-002 tiger-003	tiger-0

Table 5: Algorithms evaluated in this report.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

	Developer	Verification algorithms	Identification algorithms
89	TongYi Transportation Technology	tongyi-005	
90	Toshiba	toshiba-002 toshiba-003	toshiba-0 toshiba-1
91	Trueface.ai	trueface-000	
92	ULSee Inc	ulsee-001	
93	Veridas Digital Authentication Solutions S.L.	veridas-002	
94	Via Technologies Inc.	via-000	
95	Videonetics Technology Pvt Ltd	videonetics-001	
96	Vigilant Solutions	vigilantsolutions-006 vigilantsolutions-007	vigilantsolutions-0
97	Visidon	vd-001	vd-0
98	Vision-Box	visionbox-000 visionbox-001	
99	VisionLabs	visionlabs-006 visionlabs-007	visionlabs-7 visionlabs-008
100	Vocord	vocord-006 vocord-007	vocord-0 vocord-3
101	Winsense Co Ltd	winsense-000	
102	X-Laboratory	x-laboratory-000	
103	Xiamen Meiya Pico Information Co. Ltd	meiya-001	
104	Zhuhai Yisheng Electronics Technology	yisheng-004	yisheng-0
105	iQIYI Inc	iqface-000	
106	iSAP Solution Corporation	isap-001	

Table 6: Algorithms evaluated in this report.

4 False positive differentials in verification

False positives occur in biometric systems when samples from two individuals yield a comparison score at or above a set threshold. Most systems are configured with a threshold that is fixed for all users. False positives present a security hazard to one-to-one verification applications. They have similarly serious consequences in one-to-many identification applications. For example, in applications where subjects apply for some benefit more than once under different biographic identities e.g. visa-shopping, driving license issuance, benefits fraud, an otherwise undetected false positive might lead to various downstream consequences such a financial loss. In a surveillance application a false positive may lead to a false accusation.

This section gives empirical quantification of the variation in verification false match rates across demographics. We present results for one-to-many identification later in section 7.

We conduct several experiments with images drawn from both domestic United States and worldwide populations.

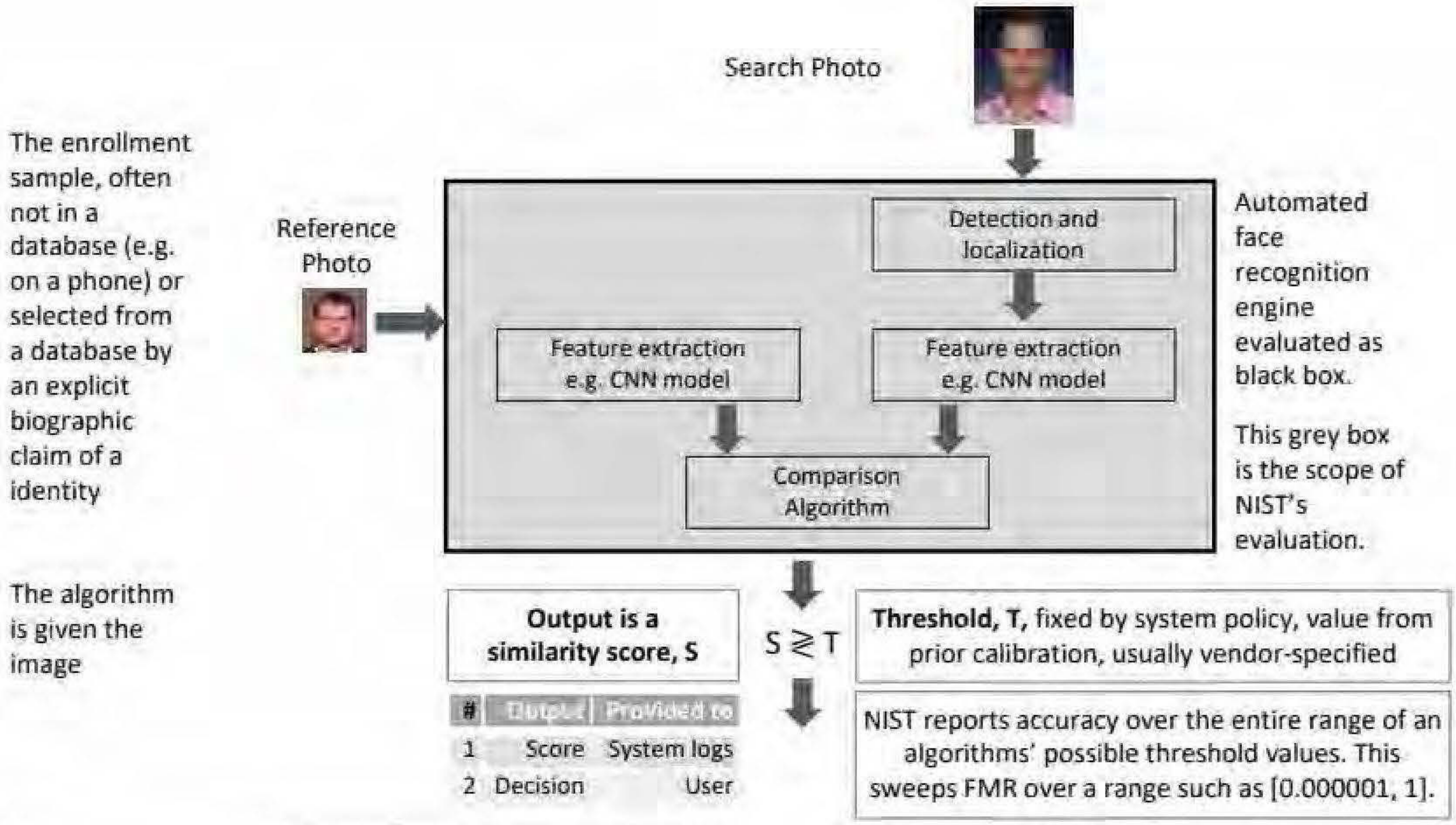
1. One-to-one application photo cross comparison, by age, sex, country-of-birth.
2. One-to-one mugshot cross comparison by age, sex, and race.
3. One-to-one visa photo cross comparison by age.

4.1 Metrics

The metrics appropriate to verification have been detailed in section 3.1. These are related to particular applications in Figure 2. The discussion in subsequent sections centers on false match rates at particular thresholds, i.e. FMR(T).

4.2 False match rates under demographic pairing

It is necessary in many biometric tests to estimate false match rates. This is done by executing imposter comparisons, and measuring false positive outcomes at some threshold(s). Historically biometric evaluations generated imposter comparisons by randomly pairing individuals, or by exhaustively comparing all individuals. As we will show in this section, this practice is inappropriate for evaluation of face recognition algorithms as it underestimates false match rates that would occur in practice. The random pairing of imposters is sometimes referred to as zero-effort pairing, mean that no effort is expended by an imposter to look like the target of the recognition attempt.



	Access Control	Non-repudiation
Role	Afford access of a person to a physical or logical resource.	Record the presence of a specific individual
Example	Door unlock. Phone unlock.	Refutation of a claim by a pharmacist that they did not dispense a particular drug, or an employer that an employee did not arrive for work.
Claim of identity	Explicit claim with an identity token such as a phone, passport or ID card.	Claim with a prior login to a system.
Threshold	High, to limit false positives	Moderate, to prevent confederates using system
Result	Acceptance decision Y/N.	Logged verification decision
Human role	Adjudicate failed rejections, to determine a false rejection, or detect an actual impostor attempt	Retrieve records to resolve a dispute
Intended human involvement frequency	Rare – approx. the false rejection rate identification rate plus prior probability of an actual mate	Rare – approx. the fraud rate multiplied by the false positive rate
Performance metric of interest	**FNMR at low FMR. See sec. 3.1, 3.2 and Tables 10, 19**	**FNMR at moderate FMR.**

Figure 2: Verification applications and relevant metrics.

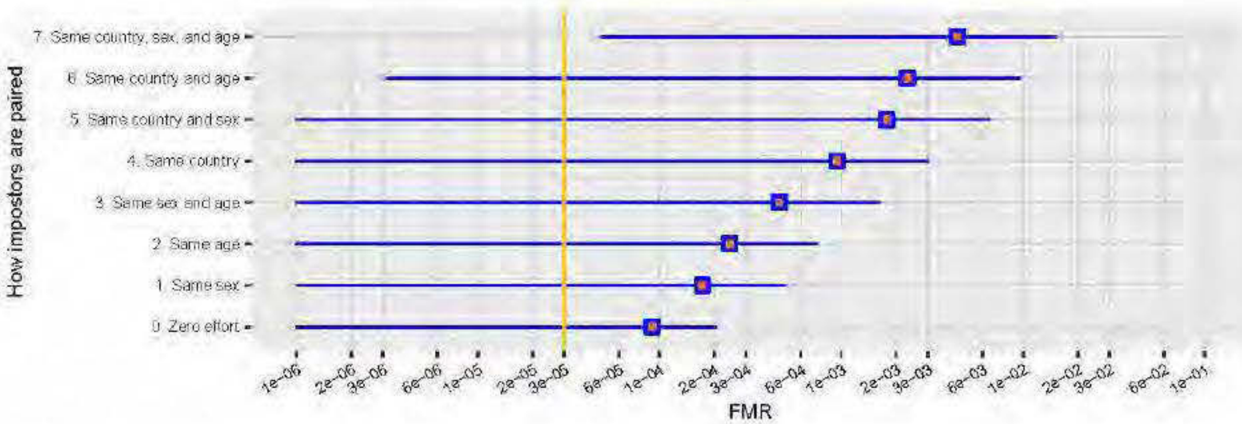


Figure 3: For application photos, the figure shows growth in one-to-one verification false match rates as the imposter demographic pairings are made more similar. At each level the point shows the mean FMR over all countries, age groups, and sexes. For example, in the second row "6. Same country and age" the mean is taken over 24 within-country times 5 within age-group times 4 within and cross sex FMR estimates, i.e. 480 FMR values. The blue line spans the 5-th to 95-th percentiles of the FMR estimates. The vertical line shows a nominal FMR value of 0.00003 obtained by setting the threshold on randomly associated i.e. zero-effort pairs of mugshot images.

Method: We used each verification algorithm to compare 442 019 application images with a disjoint set of 441 517 other application images. The two sets are subject-disjoint. The subjects were born in 24 countries. This produced 195 billion imposter scores. The images are described in Annex 2 .

The red point in the plot shows the mean of false match rates over particular sets of demographic groups.

▷ **Row 7**: The uppermost point corresponds to the mean over 240 FMR estimates, namely those comparing each of two sexes with each other, in each of five age-groups, and within each of 24 countries (2 x 5 x 24 = 240).

▷ **Row 6**: As row 7, but the average is over 480 FMR estimates that now includes different sex FMR estimates also.

▷ **Row 5**: As row 7, but now the average is over 1200 FMR estimates that additionally includes all cross-age group imposter scores.

▷ **Row 4**: As row 7, but now the average is over 2400 FMR estimates that additionally includes all cross-age and cross-sex imposter scores.

▷ **Row 3**: The average is over 5760 FMR estimates that includes 24^2 cross-country comparisons within each sex and age group.

▷ **Row 2**: The average is over 11520 FMR estimates now including different sex FMR estimates also.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

- ▷ **Row 1**: The average is over 28880 FMR estimates now including five different within-age FMR estimates also.

- ▷ **Row 0**: The average is over 57600 FMR estimates reflecting within- and between-group estimates for 24 countries, 5 age groups and 2 sexes ($24^2.5^2.2^2$).

The ordering of these rows is hand-crafted. Evaluators at DHS' Maryland Test Facility developed [20] a formal approach to showing the most influential pairing factor by quantifying information gained about FMR by having knowledge of the demographic factors, age, sex and race.

The figure shows how false match rates increase when imposters are drawn from increasingly similar demographics. This shows that fully zero-effort imposter pairings understate false match rates relative to the situation of a slightly more active imposters who would chose to present (stolen) credentials from subjects of the same sex, age and ethnicity. The practice of using zero-effort imposter pairings in tests, we think, stems from tests of fingerprint algorithm that use where friction ridge structure, particularly minutiae point arrangements, that are thought to be a developmental trait without clear genetic influence[8]

Note that our analysis has not so far documented whether particular demographic groups give higher false match rates. To address this question we introduce Figure 4 which shows results similar to those above but now for each specific country of birth.

We make the following observations:

- ▷ **Restricted pairing increases FMR**: Within each country, there is a more than order of magnitude increase in FMR between the zero-effort pair anyone-with-anyone setting, and the same-age, same-sex, same-country pairing. This re-iterates the results of the previous section, and shows it applies globally.

- ▷ **Country-of-birth matters**: For many of the different levels of demographic pairing there is between one and two orders of magnitude between the 24 countries represented in this dataset. For example when imposters are from the same sex and country but of any age, the algorithm gives FMR of 0.000046 on Polish faces and 0.0024 on Vietnamese, a fifty fold increase.

- ▷ **Regions with highest and lowest FMR**: Across algorithms often the lowest FMR is observed in Eastern European populations and the highest in East Asian populations. However there are important exceptions: Some algorithms developed in East Asia tend to give lower FMR in photos of subjects born in East Asian countries[9]. This observation and the topic of demographic differentials associated with na-

[8] Genetic influence on friction ridge structure is known: The absence of the SMARCAD1 gene leads to absence of fingerprints at birth. Further, the distance between friction ridges is smaller, on average in women than in men, and this may well be under genetic influence. The distance itself is likely not used as a biometric feature, at least not explicitly. Fingerprint pattern classes (arch, whorl etc.), however, have been shown to have regional (geographic) variations, and these were, at least historically, used in one-to-many multi-finger search strategies.

[9] See, for example, the figure in Annex 8 for algorithms from HIK, Dahua, Yitu, Alphaface, Deepsea Tencent, Toshiba.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

tional origin are covered more completely in the next section which includes results for comparison of individuals within and across national boundaries.

Discussion: The results above show that false match rates for imposter pairings in likely real-world scenarios are much higher than those from measured when imposters are paired with zero-effort. For this reason NIST has been reporting "matched-covariate" accuracy results in its FRVT evaluation of face verification algorithms [16]. Along similar lines the Australian Department of Foreign Affairs and Trade in tests it sponsors only uses same-sex imposter pairings. The effect of this is to raise thresholds, and thereby raise false non-march rates also. Thresholds increase because they are determined from non-mate scores, s, via the quantile function Q, as that value, T, which gives a proportion, FMR, at or above threshold:

$$T = Q(s, 1 - \text{FMR}) \tag{9}$$

and the set of demographically matched scores is smaller than if all possible comparisons is used.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

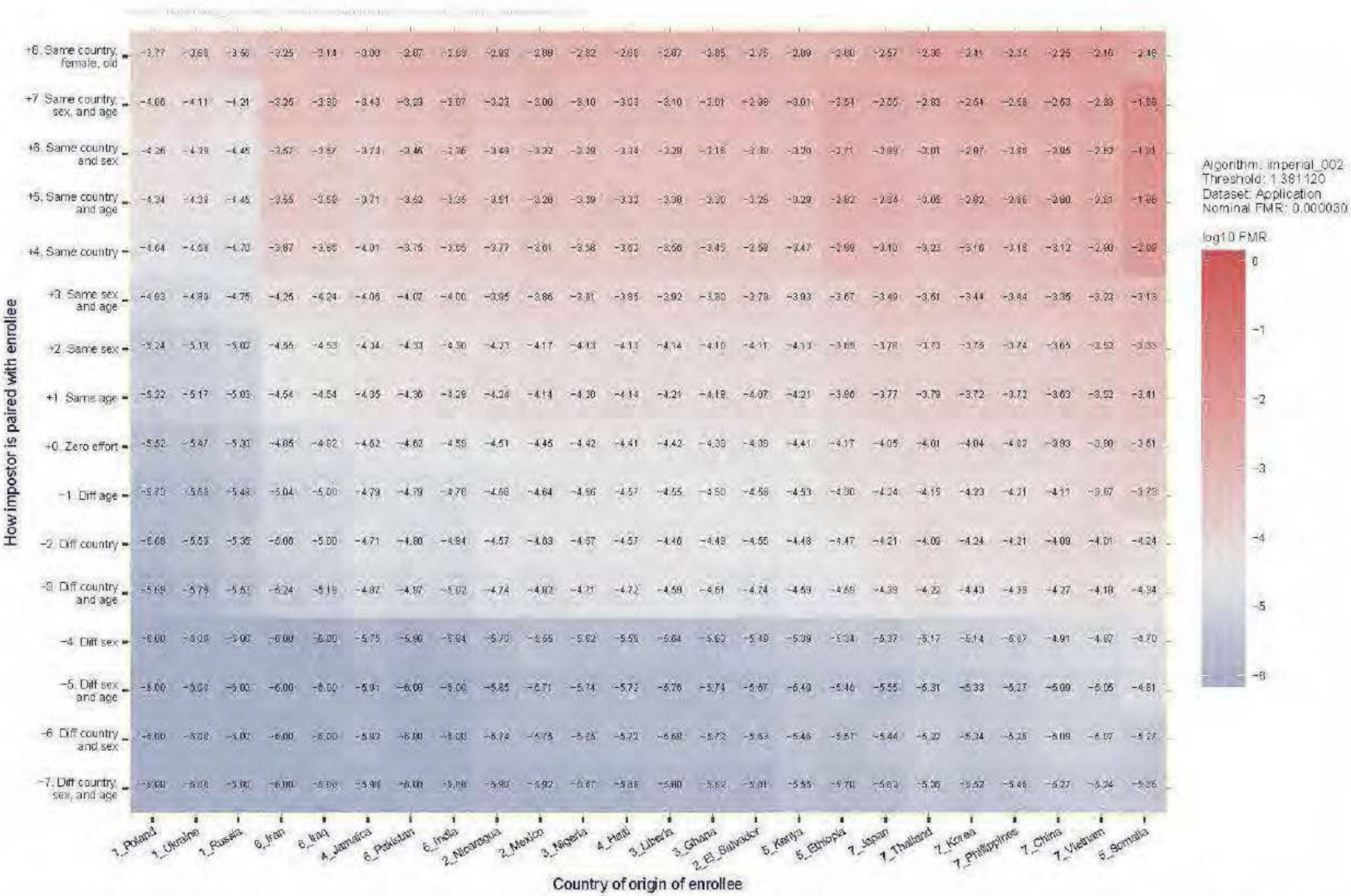


Figure 4: The heatmap shows FMR for each country-of-birth, when the imposter comparisons are drawn from increasingly demographically-matched individuals. Each cell depicts FMR on a logarithmic scale. The text value is $\log_{10}(\text{FMR})$ with large negative values encoding superior false match rates. The center row ("0. Zero effort") row compares individuals without regard to demographics. Rows above that pair imposters more closely until, in the second row, the imposters are of the same sex, age and country of origin. The top row corresponds to one particular demographic often associated with the highest FMR values. The rows below center pair for increasingly unlikely imposter pairings. For example "-5. Diff sex and age" shows FMR for imposters of different sex and age group. The countries appear in order of increasing mean FMR. Values below -6 are pinned to -6. Annex 8 contains the corresponding figure for all algorithms.

Links: Exec. Summary | Tech. Summary
False positive: Incorrect association of two subjects | False negative: Failed association of one subject
1:1 FMR | 1:1 FNMR
1:N FPIR | 1:N FNIR
T ≫ 0 → FMR, FPIR → 0 | → FNMR, FNIR → 1

4.3 False match rates within and across countries

Method: Using high quality application portraits drawn from the corpus described in Annex 2 , we compared 442 019 images from 24 countries with 441 517 images of different individuals from the same countries, yielding 195.2 billion imposter comparisons. We executed this set of comparisons with 126 verification algorithms submitted to the FRVT Verification track. These are listed in Table 4-6. We compared scores with a set of 10 thresholds to produce FMR estimates at each of those thresholds. The thresholds were computed over a set of 93 070 400 imposter comparisons made using a different set of images, namely the law enforcement mugshots detailed in Annex 1 . Each threshold was selected as the lowest value that gave FMR at or below a target FMR. The target FMR value was 0.00003.

Each photograph was assigned to the age groups defined by the intervals $(00-20]$, $(20-35]$, $(35-50]$, $(50-65]$, and $(65-99]$.

We excluded small numbers of photographs for which country of birth was not available, or for which sex was not listed as male or female.

Each comparison is accompanied by sex, country of birth and age group metadata for the two individuals represented in the photographs. Given many comparisons with the same demographic pairing, we can produce a measurement of FMR when comparing individuals from two demographic groups, for example Polish men over the age of 65 with Mexican women between 20 and 35.

Analysis: To address the issue addressed in the title of this section we produced figures depicting cross-country false match rates. Figure 5 is an example. We restricted the demographics to just men in the largest age group, $(35-50]$, and then repeated that for women. We remove sex and age from the discussion for two reasons: First, to isolate the country-of-origin effect, and, second, to reflect what real-world imposters would do: procure identity credentials from persons of the same age and sex.

Figure 5 shows cross-country FMR for one of the more accurate algorithms. Annex 7 contains corresponding figures for all algorithms, for both men and women. The annex therefore extends to more than 250 pages. We could repeat this visualization for other age groups - the results are similar. We discuss the effect of age itself later. Likewise, we could repeat the visualization for other recognition thresholds. The one adopted corresponds to a FMR = 0.00003. The trends are very similar at any threshold.

The Figure shows FMR as a heatmap. It uses a logarithmic scale, so that a FMR of 0.0001 is represented by a color and a text value of -4, i.e. log to the base 10. Low FMR values are shown in blue. High FMR values are shown in red. A grey color connotes the target FMR value ($\log_{10} 0.00003 = -4.5$). High FMR values present a security concern in verification applications.

Discussion: From the Figure and those in the annexes, we make a number of observations. First by assigning

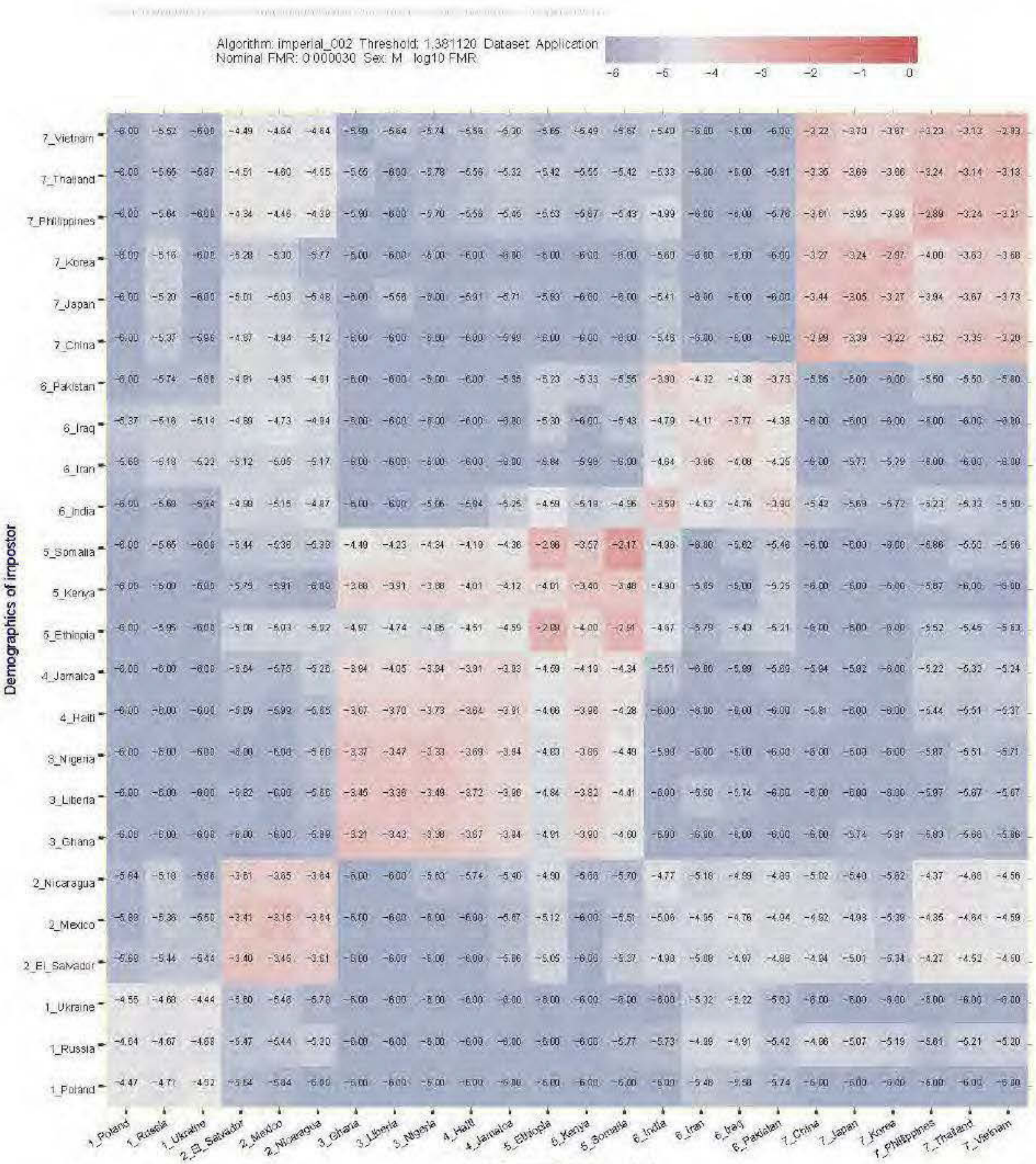


Figure 5: For 24 countries in seven regions the figure shows false positive rates when the reference algorithm is used to compare single photos of mid-aged male subjects from the countries identified in the respective columns. The threshold is to a preset fixed value everywhere. Each cell depicts FMR on a logarithmic scale. The text value is $\log_{10}(\text{FMR})$ with large negative values encoding superior false match rates. Annex 7 contains the corresponding figure for all algorithms.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

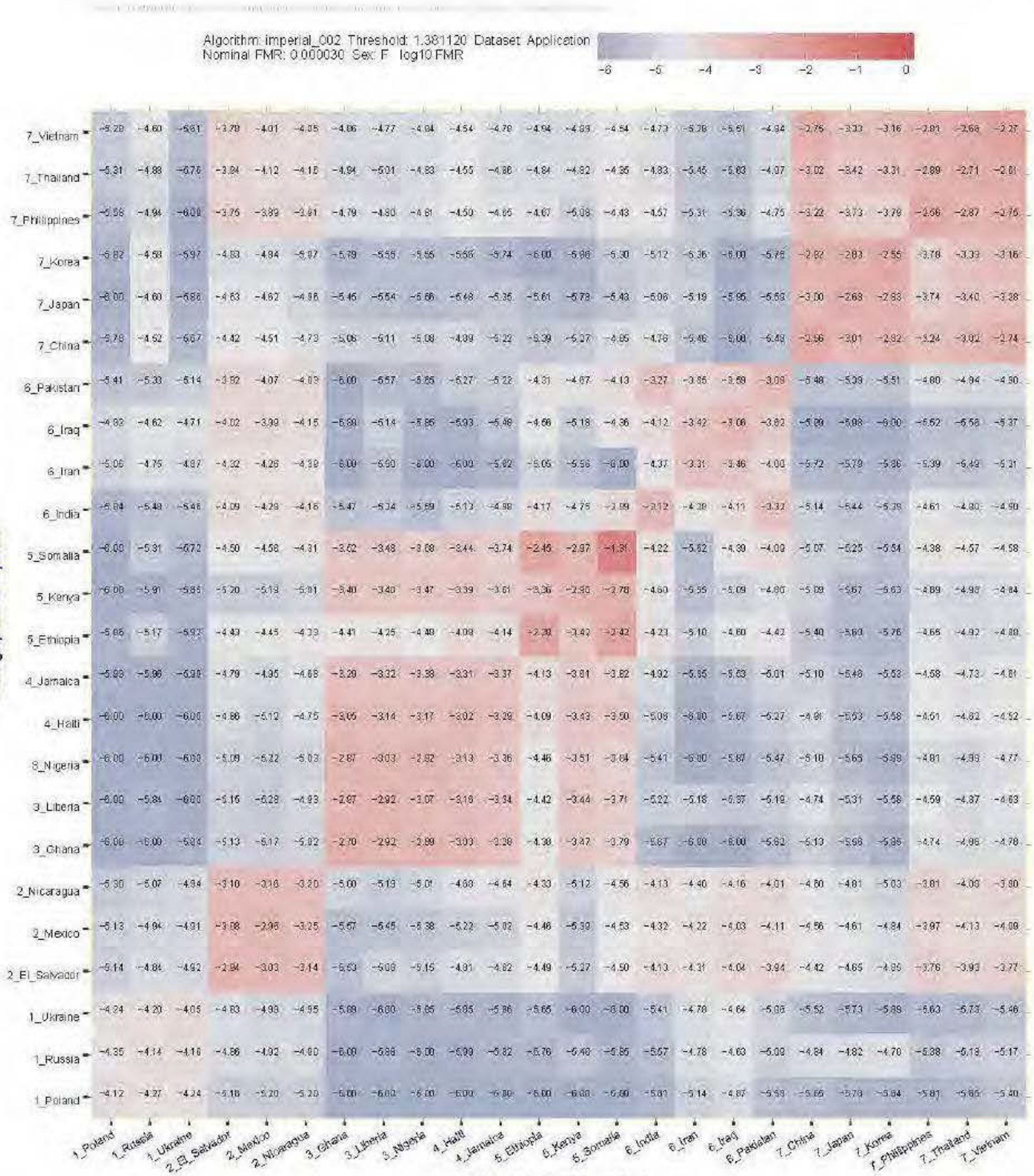


Figure 6: For 24 countries in seven regions the figure shows false positive rates when the reference algorithm is used to compare single photos of mid-aged female subjects from the countries identified in the respective columns. The threshold is to a preset fixed value everywhere. Each cell depicts FMR on a logarithmic scale. The text value is $\log_{10}(\text{FMR})$ with large negative values encoding superior false match rates. Annex 7 contains the corresponding figure for all algorithms.

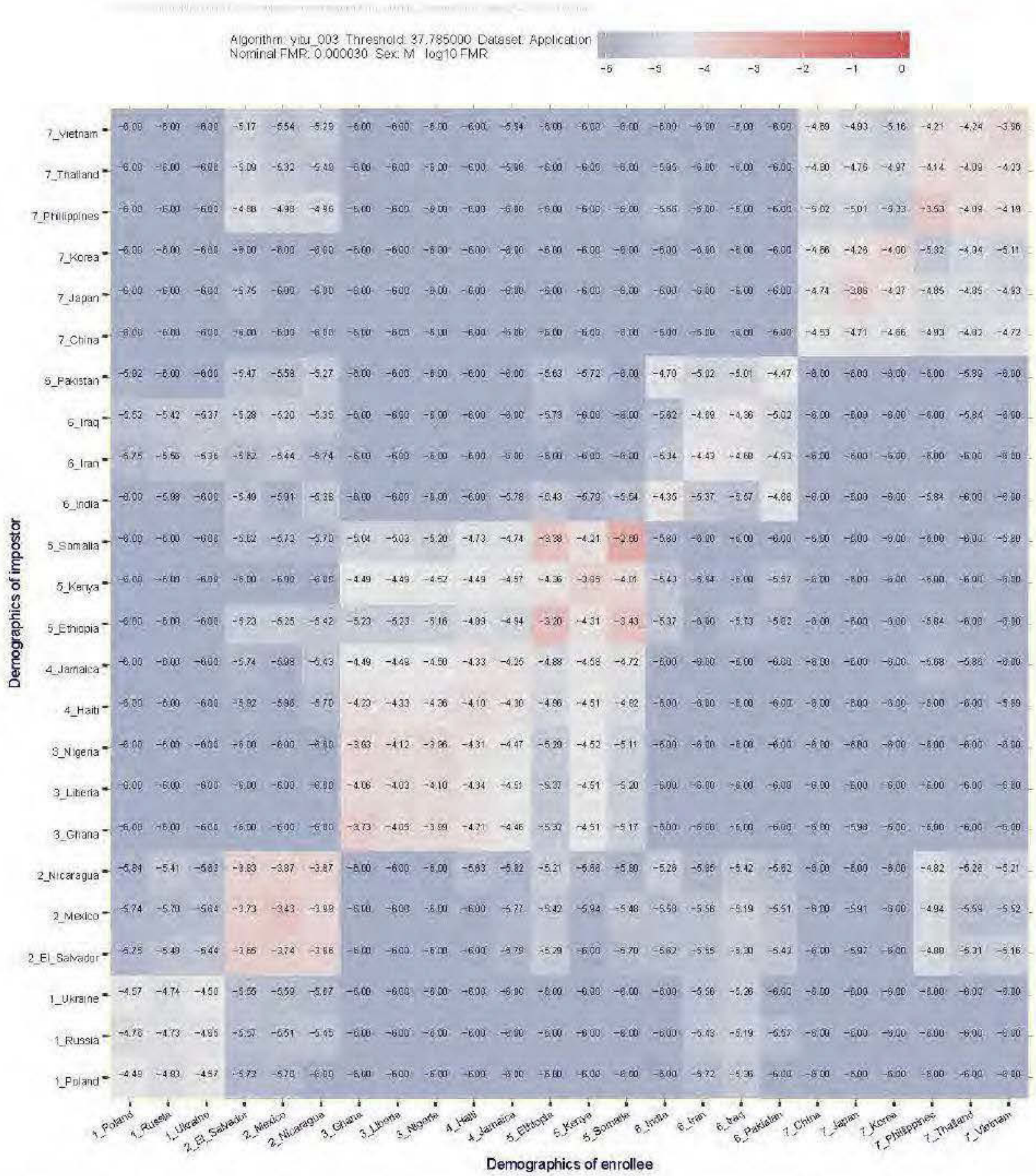


Figure 7: For 24 countries in seven regions the figure shows false positive rates when the **Chinese-developed** *algorithm is used to compare single photos of mid-aged male subjects from the countries identified in the respective columns. The threshold is to a preset fixed value everywhere. Each cell depicts FMR on a logarithmic scale. The text value is* $\log_{10}(\text{FMR})$ *with large negative values encoding superior false match rates.* Annex 7 *contains the corresponding figure for all algorithms.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

countries into the following regions

- ▷ 1: Eastern Europe - Russia, Poland and Ukraine
- ▷ 2: Central America - Mexico, Honduras, El Salvador, Nicaragua
- ▷ 3: West Africa - Ghana, Liberia, Nigeria
- ▷ 4: The Caribbean - Haiti, Jamaica
- ▷ 5: East Africa - Ethiopia, Kenya, Somalia
- ▷ 6: South Asia - India, Iran, Iraq, Pakistan
- ▷ 7: East Asia - China, Japan, Korea, Philippines, Thailand, Vietnam

we see a block structure, in particular a block-diagonal structure indicative of strongly correlated false match rates within region. For example it is true that when comparing photos of individuals from East Africa with those from Eastern Europe, most algorithms give very low FMR. The more interesting results are within-region, around the diagonal, and between regions along the diagonal. We now note the following common trends, and then some notable exceptions. We then conclude with some comments on what the ideal situation would be, and on the meaning. Each Annex includes a "contact sheet" which shows all heatmaps on a single page as thumbnails. The idea is to show macroscopic behavior across all algorithms. When viewed on a computer the figure has very high resolution and zooming in reveals full detail; when printed it will likely just show coarse trends.

- ▷ Nominal FMR in Eastern Europe: For many algorithms, FMR within Eastern Europe is close to the nominal target false match rate i.e. a grey color, $-5 \leq \log_{10} \text{FMR} \leq -4$. There are few exceptions to this, even for algorithms developed in China, Western Europe and the USA.
- ▷ Higher FMR in East Africa: For almost all algorithms the highest FMR is for comparison of Somali faces. We suspected this could be due to mislabeled data or statistical (in)significance but rejected those possibilities[10]. Further the FMR is high within Ethiopia and between Ethiopia and Somalia. Similarly Kenya-Kenya comparisons give high FMR, although somewhat reduced. In a substantial majority of photos of Somalian women, the subject is wearing full head dress that typically covers the hair and ears leaving only the face exposed. While this might produce false positives, headwear is almost always absent in photographs of men. Further work is needed to explain the observation in more detail.

[10] We discount that this result is anomalous as follows: 1. The sample size may be small for this study, but not absolutely small: The Somalia-Somalia FMR measurement is obtained from 1 733 116 comparisons involving 2632 images of 1974 males. 2. The effect persists when comparing Somalian and Ethiopian faces, and we'd suspect that ground-truth labelling errors - instances of one person being present two IDs - would not persist across national boundaries. 3. In addition to high FMR, which is a count of high imposter scores, the mean similarity score is also very high, an observation that again applies to all algorithms.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

▷ Higher FMR in West Africa too: The countries with the second highest FMR tend to be in West Africa, i.e. Ghana, Liberia and Nigeria. These countries do not share any borders. The high FMR values occur almost equally within and between countries.

▷ Higher FMR between West Africa and the Caribbean: Elevated FMR occurs when comparing faces of individuals from countries in West Africa with those in the Caribbean.

▷ Higher FMR between West and East Africa: Elevated FMR occurs when comparing faces of individuals from countries in West Africa and Kenya. The effect is often lower than within either region alone. However, the high FMR does not extend to comparisons of West African and Ethiopian or Somali faces.

▷ Higher FMR in East Asia: It is very common for algorithms to give high FMR within East Asian countries and between them. For the algorithm shown, Vietnamese faces strongly match other Vietnamese, and with all the other countries in the region. The East Asian block often divides into northern and southern blocks with reduced, but still high, FMR when individuals are compared between those blocks (e.g. Korea and Vietnam).

▷ Some Chinese algorithms give nominal FMR when comparing Chinese: As shown in Annex 7 some algorithms developed in China exhibit much reduced FMR on the East Asian population - for example, see Figure 7. These algorithms are from Megvii, Meiya, Hik Vision, Dahua, X-Laboratory, Yitu and SHU (Shanghai University Film Academy). For Deepsea Tencent the same applies, but less prominently in South East Asia. In some cases the effect is only apparent for comparisons involving images of Chinese, e.g. Star Hybrid. Other Chinese algorithms, however, exhibit the more common trend of producing elevated FMR across East Asia. These include developers of more accurate algorithm such as Alphaface, Deepglint and Sensetime. Thus it is not sufficient for an algorithm to be developed in China for it to mitigate the FMR increase on images from the local population.

▷ One of the most accurate algorithms produces more uniform FMR: The corresponding Figure for the Yitu-003 algorithm - Figure 7) - shows that the demographic differentials in FMR are attenuated. As noted the FMR values for comparisons within East Asia are near the nominal value. Notably, however, this applies to West Africa also. This appears to be an important result, as it is a proof that some algorithms do not exhibit higher FMR in those populations. Yitu reported in a meeting in London in October 2017 that its training data included on order of 10^9 photographs of an unspecified (lower) number of Chinese nationals. Whether that is the entirety of their training data is not known.

▷ Developer dependency does not apply to South Asia: Neither Lookman nor Tiger IT's algorithm produce nominal FMR on the S. Asian imposter comparisons.

- ▷ Magnitudes are large: The East African FMR values are often two orders of magnitude higher than the nominal value and those recorded within Eastern Europe. That is, the $\log_{10}$ FMR values are +2 higher corresponding to FMR that is of order 100 times larger than the de-facto baseline. From a security perspective this is analogous to using a two-digit PIN instead of the common four digits. For West Africa, the FMR values are between one and two orders of magnitude above baseline. A shift of 1.4 on the logarithmic scale corresponds to a factor of 25 increase, for example.

- ▷ Anomalies in the figures: The cross-country heatmaps for the SIAT-004, Panasonic PSL-001, and Sensetime-002 algorithms are mostly red, indicating high false match rates for all comparisons. This may arise because the threshold used was computed over comparisons of a different kind of images - mugshots not application portraits. The algorithms are told what kind of image they are being given at the time features are extracted from the image. The consequence is that the imposter distribution for mugshots looks different to that for the application images, and thus thresholds are not portable. This would present an operational issue to any end-user not informed to set the threshold accordingly. In any case, while the heatmaps are mostly red, they still exhibit the same kind of FMR variations seen for many other algorithms.

Discussion: The heatmap figures of Annex 7 show a widespread elevation of false match rates in African faces relative to those in Eastern Europe. The reasons for these shifts are unknown. We did not make any attempts to explain the effects. To summarize the effect we include the scatter plots of Figures 10 - 9. Each point corresponds to one algorithm. Its coordinates show false match rates within West Africa against those within Eastern Europe. The degree to which the point is above the diagonal line shows the extent that FMR in the African countries exceeds that in the Eastern European ones.

We note several outcomes of this visualization.

- ▷ **Worst case** In the scatter plot for African women Figure 9 there is a cluster of algorithms located near $x = 0.00012$ and $y = 0.003$. Compared to the target FMR value of 0.00003 (the vertical line) there is a near four-fold increase in FMR of women over men. Much more significantly there is a more than 100-fold vertical excursion from white men to African women.

- ▷ **Dispersion** Some algorithms, most notably those from Sensetime give FMR much different to the target value. The threshold was set using Annex 1 mugshots but the Figure reflects FMR measured over comparison of Annex 2 application photos. Both sets of photos are well illuminated portraits, so this instability across datasets would be unwelcome, especially if an algorithm were to be fielded on imagery qualitatively different. Many algorithms do give the expected FMR for white men FMR = 0.00003 as seen in Figure 8.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

Figures 10 and 11 repeat the scatterplot summaries for the East Asian demographic too. The picture there is more interesting. While the same pattern is present, it is clear that some algorithms developed in China do not give elevated false match rates relative to Eastern Europeans. The absence of the effect is important in that it implies high FMR in that population is not inevitable. We did not see a corresponding improvement for South Asian faces for the few algorithms we understand were submitted by developers there (in India and Bangladesh).

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

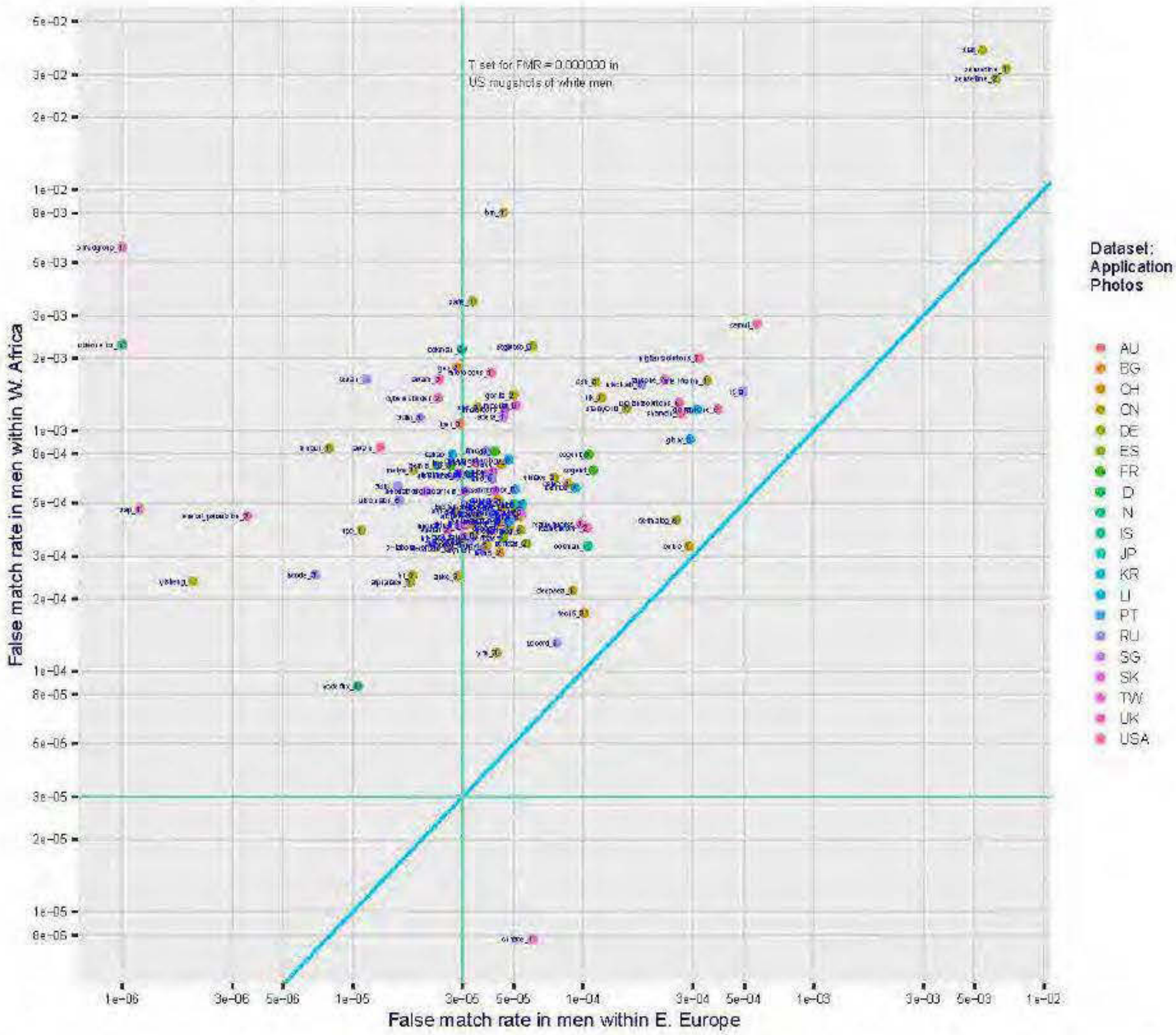


Figure 8: The scatter plot shows FMR when comparing same-age men within and across three Eastern European countries (Russia, Ukraine, Poland), against FMR obtained comparing men within and across three West African countries (Ghana, Liberia, Nigeria). The threshold is fixed for each algorithm to give the FMR noted in the annotation over white men in the U.S. mugshot database. This is indicated by the vertical and horizontal green lines. The blue diagonal line $y = x$ is included to show "over/under". The color code identifies the domicile of the developer - some multinationals conduct research elsewhere. Training data likewise may originate elsewhere.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

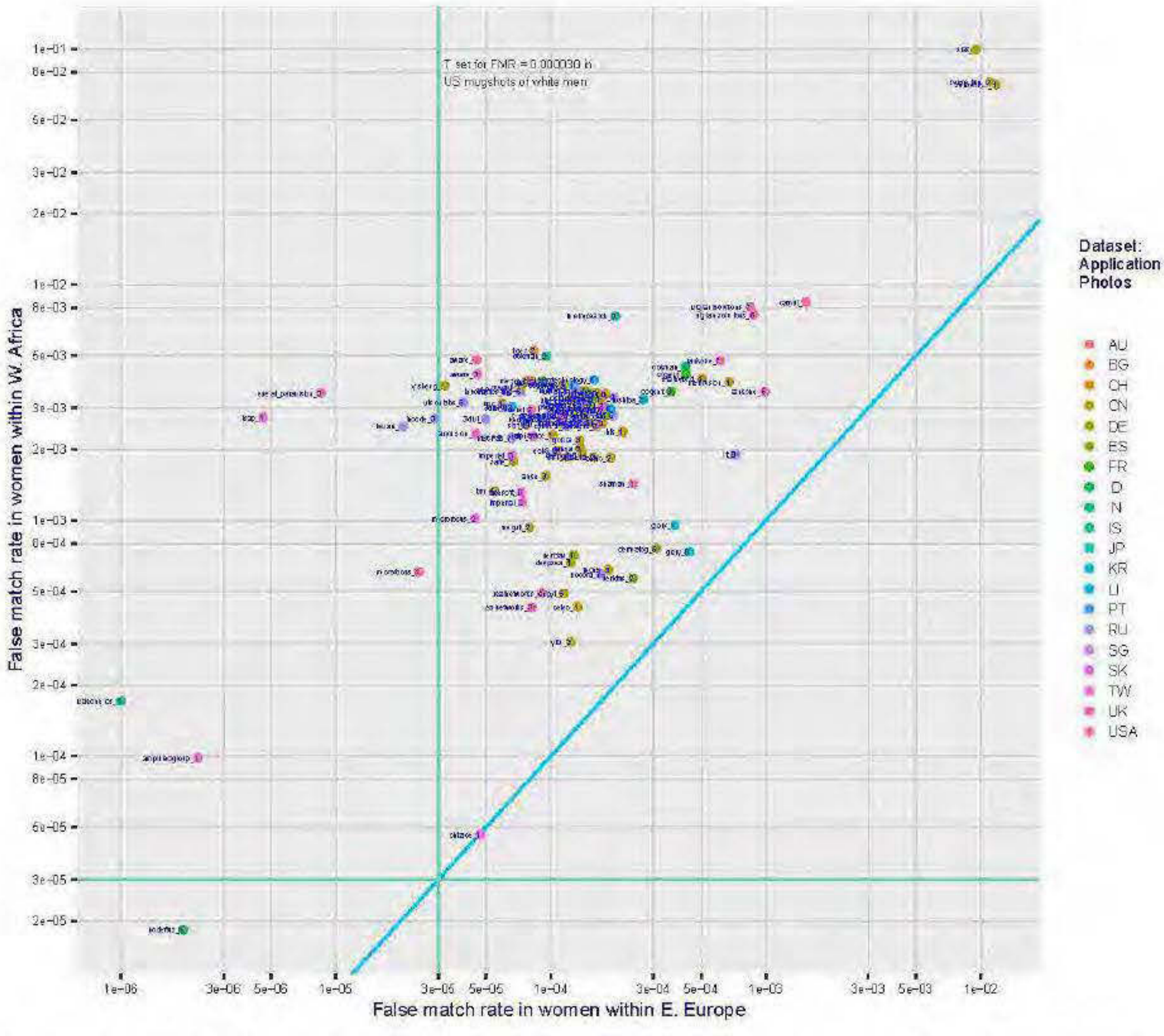


Figure 9: The scatter plot shows FMR when comparing same-age **women** *within and across three Eastern European countries (Russia, Ukraine, Poland), against FMR obtained comparing women within and across three West African countries (Ghana, Liberia, Nigeria). The threshold is fixed for each algorithm to give the FMR noted in the annotation over white men in the U.S. mugshot database. This is indicated by the vertical and horizontal green lines. The blue diagonal line $y = x$ is included to show "over/under". The color code identifies the domicile of the developer - some multinationals conduct research elsewhere. Training data likewise may originate elsewhere.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

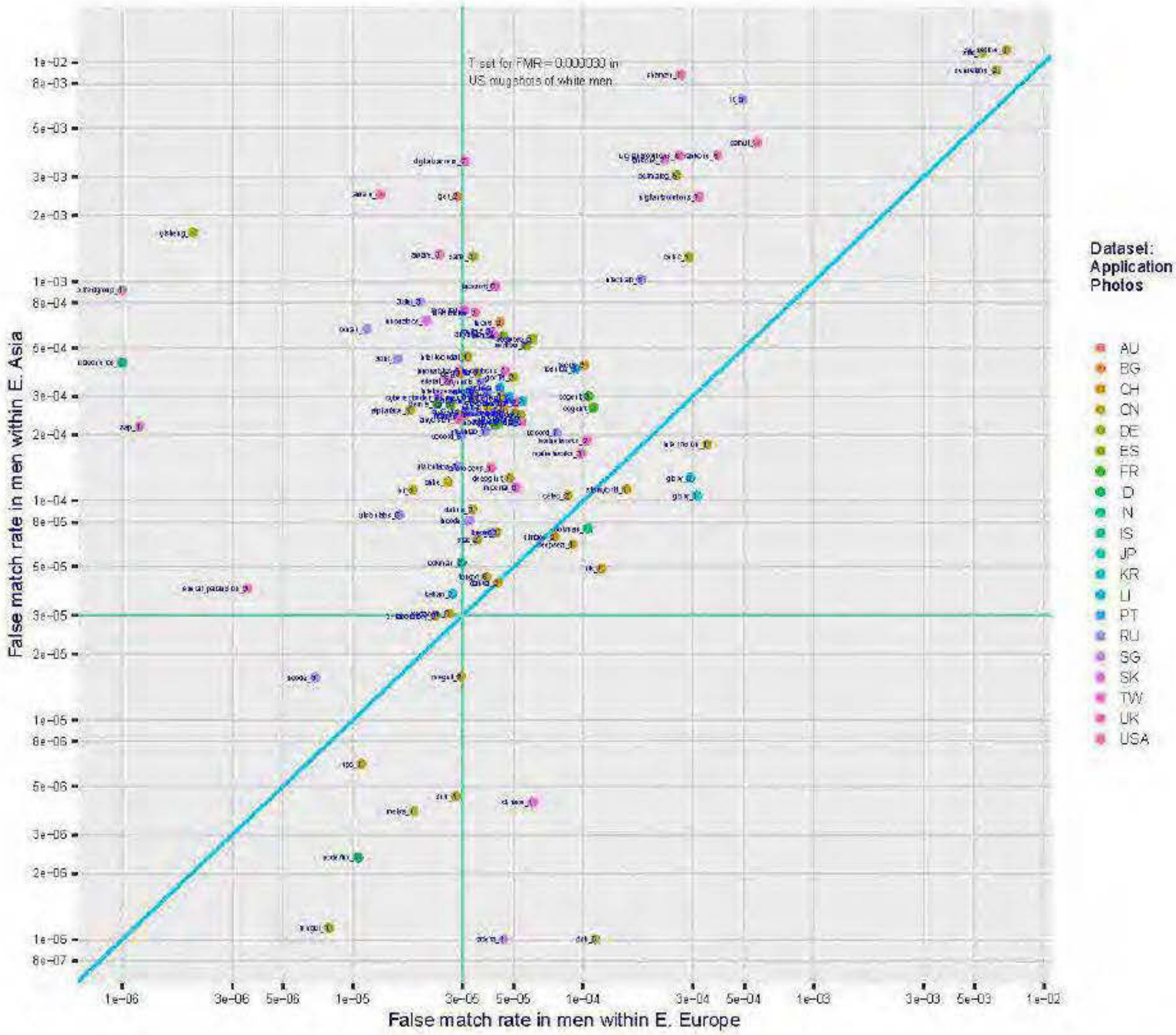


Figure 10: The scatter plot shows FMR when comparing same-age men *within and across three Eastern European countries (Poland, Russia, Ukraine), against FMR obtained comparing men within and across six East Asian countries (China, Japan, Korea, Philippines, Thailand and Vietnam). The threshold is fixed for each algorithm to give the FMR noted in the annotation over white men in the U.S. mugshot database. This is indicated by the vertical and horizontal green lines. The blue diagonal line $y = x$ is included to show "over/under". The color code identifies the domicile of the developer - some multinationals conduct research elsewhere. Training data likewise may originate elsewhere.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

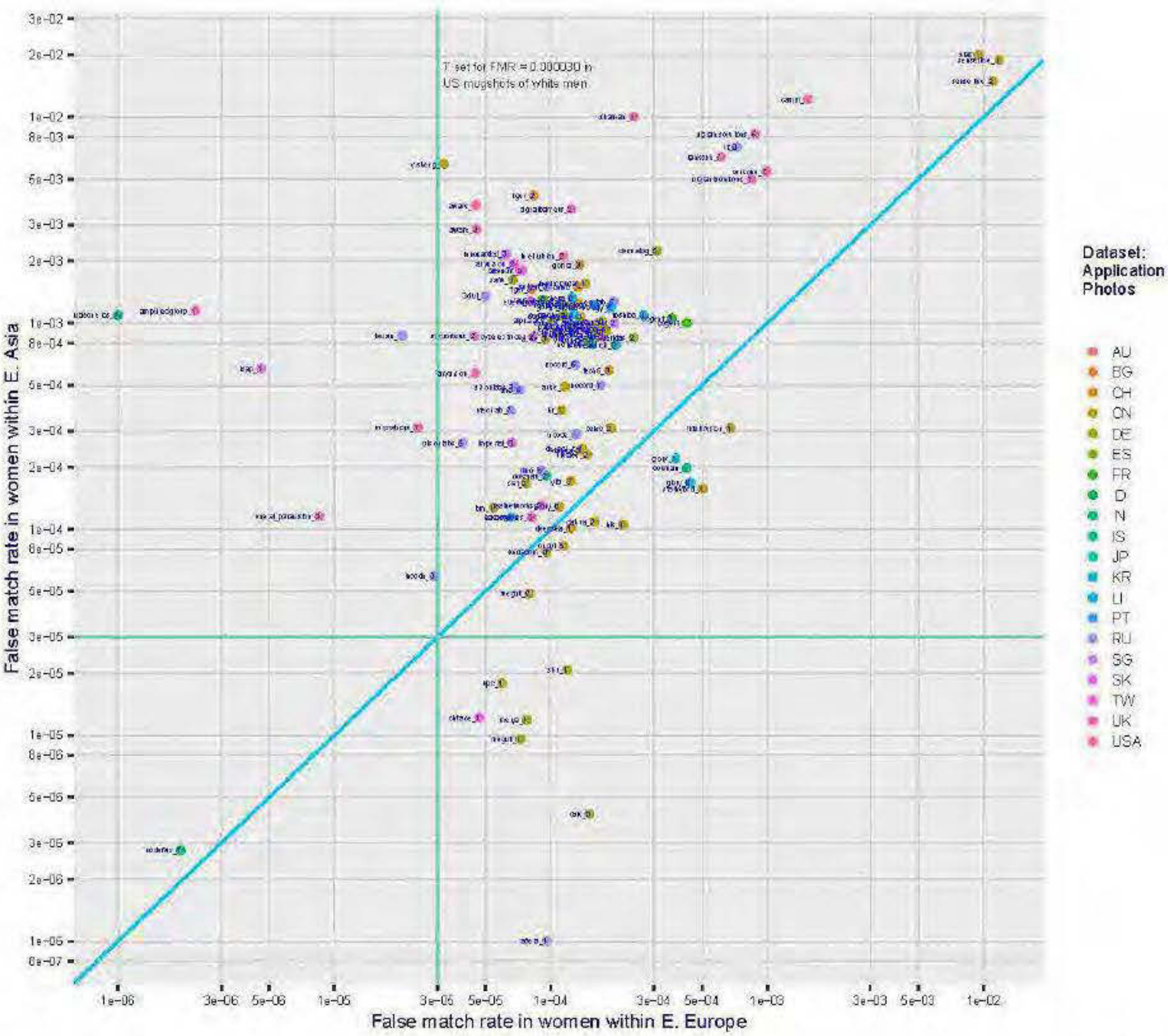


Figure 11: The scatter plot shows FMR when comparing same-age **women** *within and across three Eastern European countries (Poland, Russia, Ukraine), against FMR obtained comparing women within and across six East Asian countries (China, Japan, Korea, Philippines, Thailand and Vietnam). The threshold is fixed for each algorithm to give the FMR noted in the annotation over white men in the U.S. mugshot database. This is indicated by the vertical and horizontal green lines. The blue diagonal line $y = x$ is included to show "over/under". The color code identifies the domicile of the developer - some multinationals conduct research elsewhere. Training data likewise may originate elsewhere.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

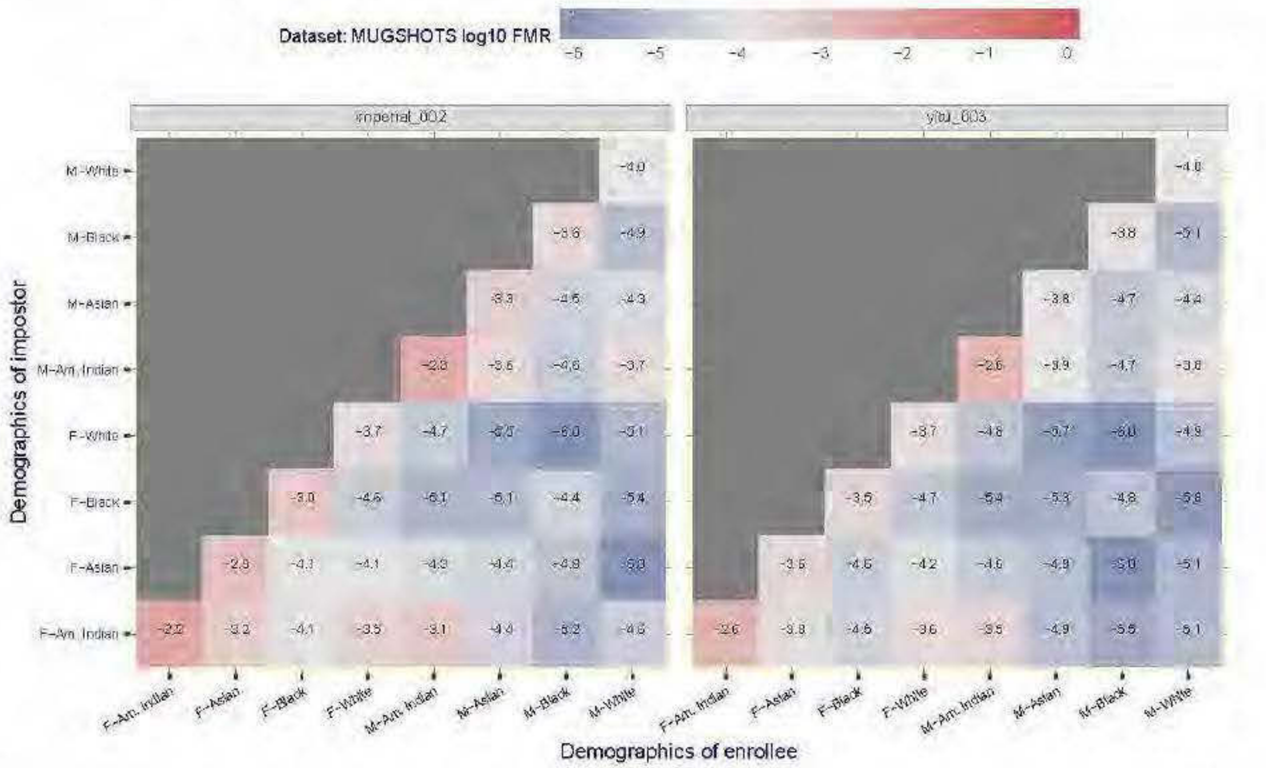


Figure 12: For mugshot photos tagged with one of four race labels and a sex label, the heatmaps show false positive rates for comparison of randomly selected photos from the groups identified in the respective rows and columns. Two algorithms are used, one in each panel, and the threshold for each is set to a fixed value everywhere. The value is the smallest threshold that gives FMR $\leq$ 0.0001 on the white male imposters. Each cell depicts FMR on a logarithmic scale. The text value is $\log_{10}$(FMR) with large negative values encoding superior false match rates. Annex 6 contains the corresponding figure for all algorithms.

4.4 Dependence of FMR on race in United States mugshots

Method: Using high quality mugshot portraits from the mugshot images detailed Annex 1 , we apply each verification algorithm to conduct 3 million comparisons for each of the eight demographics defined by two sexes and four races. The origin and meaning of these labels is described in the Annex. We executed this set of comparisons with 126 verification algorithms submitted to the FRVT Verification track. These are listed in Tables 4-6. We compared scores with a threshold to produce FMR estimates for each demographic pairing. Each threshold was selected as the lowest value that gave FMR at or below a target FMR. The target FMR value was 0.0001. The threshold was computed over the set of 3 000 000 mugshot imposter comparisons made for white males. Thus, by design, the FMR for that demographic is exactly 0.0001.

We excluded photographs for which race or sex was unavailable or unknown. We did not report comparisons by age-group.

Analysis: As with the international set of application photos, we use the heatmap to show cross-demographic

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

false match rates, including cross-sex. Heatmaps for two algorithms are shown in Figure 12. The Figure shows FMR as a heatmap. It uses a logarithmic scale, so that a FMR of 0.0001 is represented by a color and a text value of -4, i.e. log to the base 10. Low FMR values are shown in blue. High FMR values are shown in red. A grey color connotes the target FMR value ($\log_{10} 0.0001 = -4$). High FMR values present a security concern in verification applications. Corresponding figures for all algorithms appear in Annex 6

Figure 13 extracts the within-sex and within-race diagonal elements of those figures and summarizes the results for all algorithms, ordering the result by worst-case FMR elevation.

Discussion: From the figure, and those in the annex, we make a number of observations.

- ▷ Higher FMR in women: As with application photos, most algorithms give systematically higher false match rates in women than in men. The magnitude of this difference is lower with mugshots than with application photos.
- ▷ Highest FMR in American Indians: First, the highest FMR occurs in images of American Indians[11]. For the Imperial-002 algorithm featured in Figure 12 the FMR for American Indian women is 0.0068, i.e. a 68 fold increase over the FMR of 0.0001 in white males. In men, the multiple is 47. Why such large increases occur is not known. One component of the increase may stem from database identity labelling errors[12]. We discount this possibility because the database has otherwise excellent ground-truth integrity, supported by fingerprint enrollments.
- ▷ Higher FMR in Asian and Black women: There are order-of-magnitude increases in FMR in mugshots of Asian and Black women. Some algorithms developed in China reduce this differential, for example Yitu-003 in the right panel of Figure 12.

[11] The data supplied to NIST tags this group with letter "I" per the EBTS standard which describes this group as "American Indian, Eskimo, Alaskan native, or a person having origins in any of the 48 contiguous states of the United States or Alaska who maintains cultural identification through tribal affiliation or community recognition". In the figures we replace the letter "I" with "American Indian" to distinguish from subjects from India in the international datasets.

[12] Specifically instances of "one person under two IDs" can cause apparent false positives, that are actually true positives.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

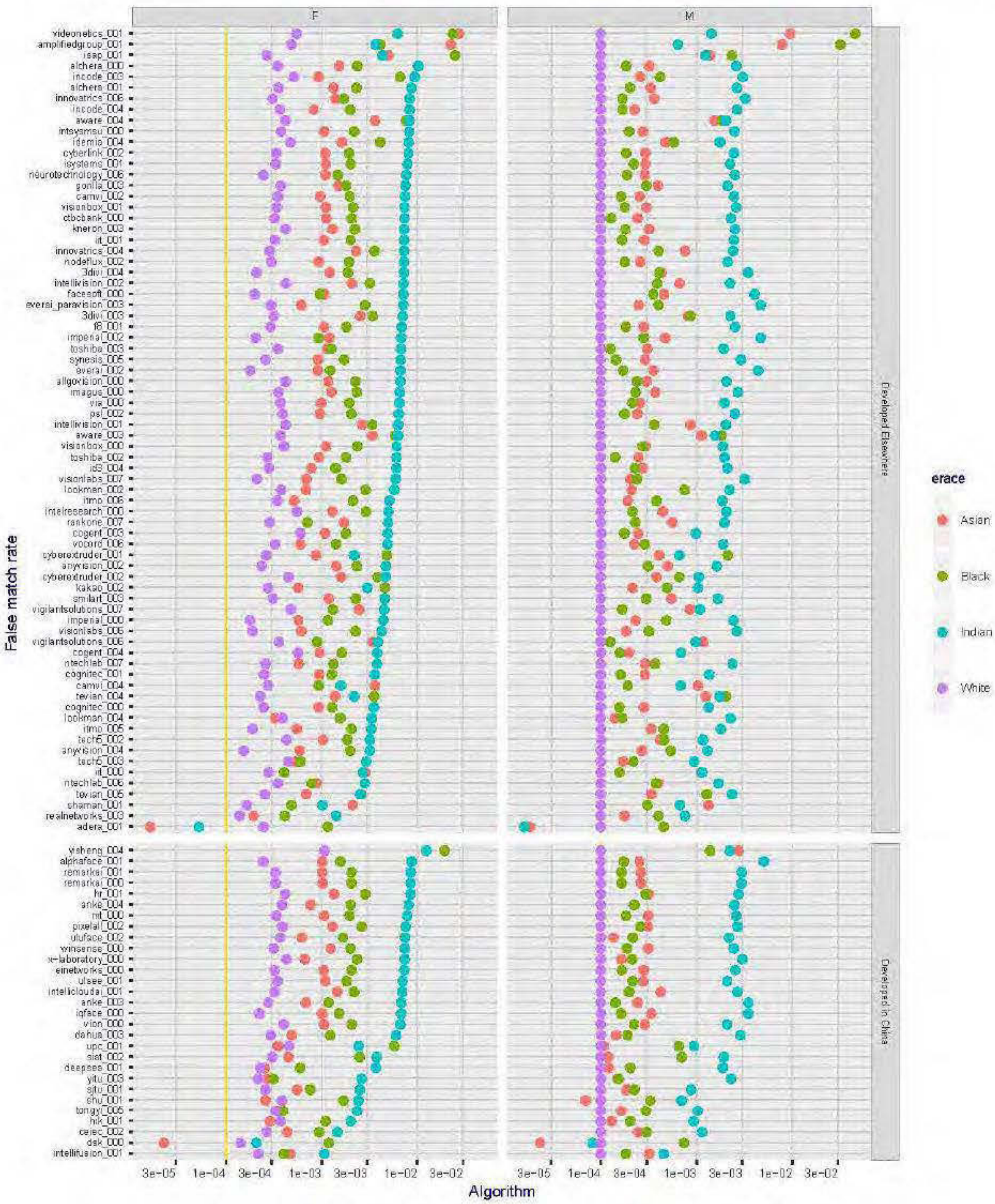


Figure 13: For each verification algorithm, the dots give the false match rates for same-sex and same-race imposter comparisons. The threshold is set for each algorithm to give FMR = 0.0001 on white males (the purple dots in the right hand panel). The algorithms are sorted in order of worst case FMR, usually for American Indian women. Algorithms developed in China appear in the lower panel.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

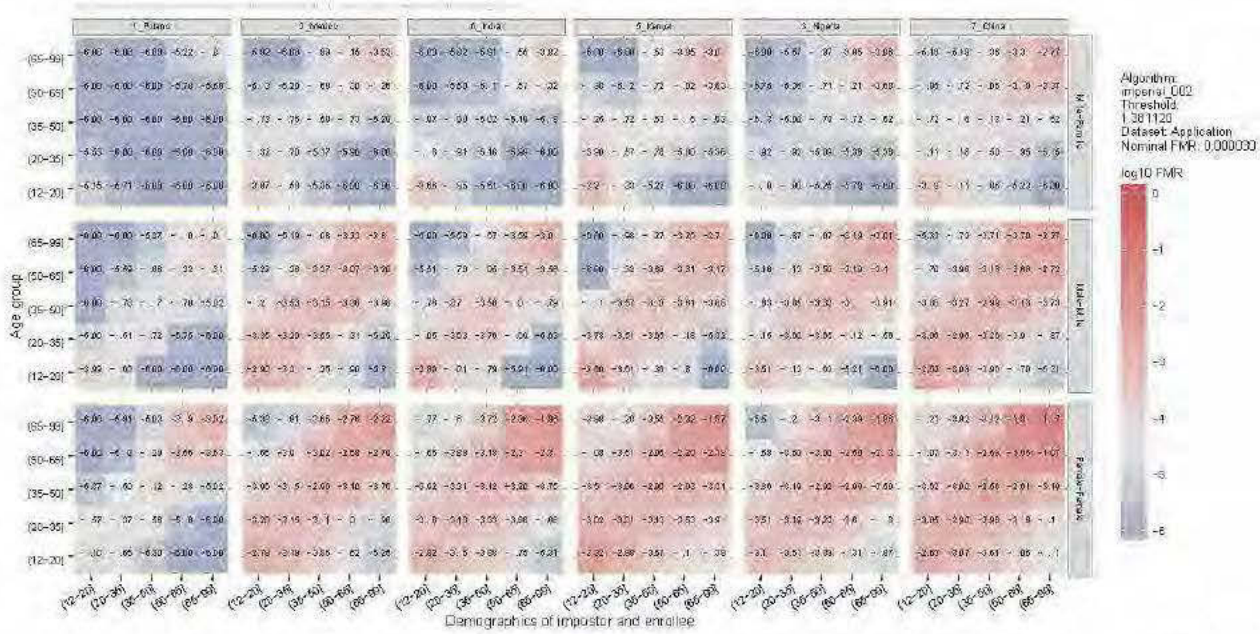


Figure 14: For six countries selected for the high number of images in the dataset and from distinct regions the heatmaps show cross-age false match rates for imposters of the same sex from the age groups given on the respective axes. Each cell depicts FMR on a logarithmic scale. The text value is $\log_{10}(\text{FMR})$ with large negative values encoding superior false match rates. Annex 9 contains the corresponding figure for all algorithms.

4.5 Do some or all algorithms yield more false positives on certain age groups

Method: Using high quality application portraits drawn from the corpus described in Annex 2 , we compared 442 019 images from 24 countries with 441 517 images of different individuals within and across age groups $(00-20], (20-35], (35-50], (50-65]$, and $(65-99]$.

We executed this set of comparisons with 126 verification algorithms submitted to the FRVT Verification track. These are listed in Tables 4-6. Each comparison yield a score. When many scores are compared with a fixed threshold, we obtain an estimate of the false match rate. The threshold was computed over a set of 93 070 400 imposter comparisons made using a different set of images, namely the mugshots detailed in Annex 1 . The threshold is the smallest value that for which the FMR is less than or equal to 0.00003. This was repeated for other thresholds giving FMR $\{0.000001, 0.000003, 0.00001, 0.00003, 0.0001, 0.0003, 0.001, 0.003, 0.01, 0.03\}$.

Each comparison is accompanied by sex, country of birth and age group metadata for the two individuals represented in the photographs. We excluded small numbers of photographs for which age information was unavailable or for which sex was not listed as male or female.

Given many comparisons with the same demographic pairing, we can produce a measurement of FMR when comparing individuals from two age groups, for example Polish men over the age of 65 with Polish men under 20.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

Figure 15: For visa photos from all countries, the heatmap shows for one algorithm cross-age false match rates for imposters of the same sex. Each cell depicts FMR on a logarithmic scale. The text value is $\log_{10}(\text{FMR})$ with large negative values encoding superior false match rates. The threshold is fixed to the value that gives a FMR of 0.0001 over all zero-effort imposter pairs. Annex 10 contains the corresponding figure for all algorithms.

Analysis: To address the issue of age we produced figures depicting cross-age false match rates. We do this within-country only, as cross-country effects have been covered in section 4.3. We include male-male, female-female, and also male-female comparisons (although they are of less interest operationally). Figure 14 is an example, showing results for one of the more accurate algorithms. The Figure includes results for six countries, one per region. We dropped one region (the Caribbean) and 18 of the 24 countries because the effects are similar everywhere.

Figure 14 shows cross-age group FMR for one of the more accurate algorithms. Annex 9 contains corresponding Figures for all algorithms, and therefore extends to more than 130 pages.

Discussion: From Figure 14 and those in the annex, we make these observations.

- ▷ Lower FMR for persons in different groups: In almost all cases - for all algorithms, countries of origin and both sexes, comparison of images of persons in different age groups yields lower (better) false match rates than for persons in the same age group. This, obviously, is an aggregate result; it will generally be possible to find some individuals from different age groups who produce high imposter scores but this will be increasingly difficult as the age difference increases.

- ▷ Highest FMR in the oldest age group: For women from all most countries, comparison of images of individuals in the 65-and-over age group produce the highest false match rates. For men this is often true also.

- ▷ High FMR in the youngest age group: For both sexes, but men in particular, comparison of images of persons in the 12 – 20 age group produce high false match rates. The dataset does not include any subjects below 12. Below that age we consider a smaller dataset of visa photographs (see Annex 3) that includes individuals in age groups $(0, 4]$ and $(4, 10]$. The results are included in the heatmap of Figure 15. Note that each FMR estimate is formed from comparisons from all countries, not just one, so they hide the geographic idiosyncrasies of the algorithms.

  These results are similar to those reported by Michalski et al. [28] for false positives in children using one commercial algorithm. The report also shows false negative ageing effects broken out by age at enrolment, and time lapse.

- ▷ Lower FMR across sex: Comparison of images of persons of different sex usually produces very low FMR. However, within the youngest and oldest age groups, FMR is again higher and substantially above the nominal FMR.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

Links: EXEC. SUMMARY TECH. SUMMARY
False positive: Incorrect association of two subjects
False negative: Failed association of one subject
1:1 FMR 1:1 FNMR 1:N FPIR 1:N FNIR
T ≫ 0 → FMR, FPIR → 0
→ FNMR, FNIR → 1

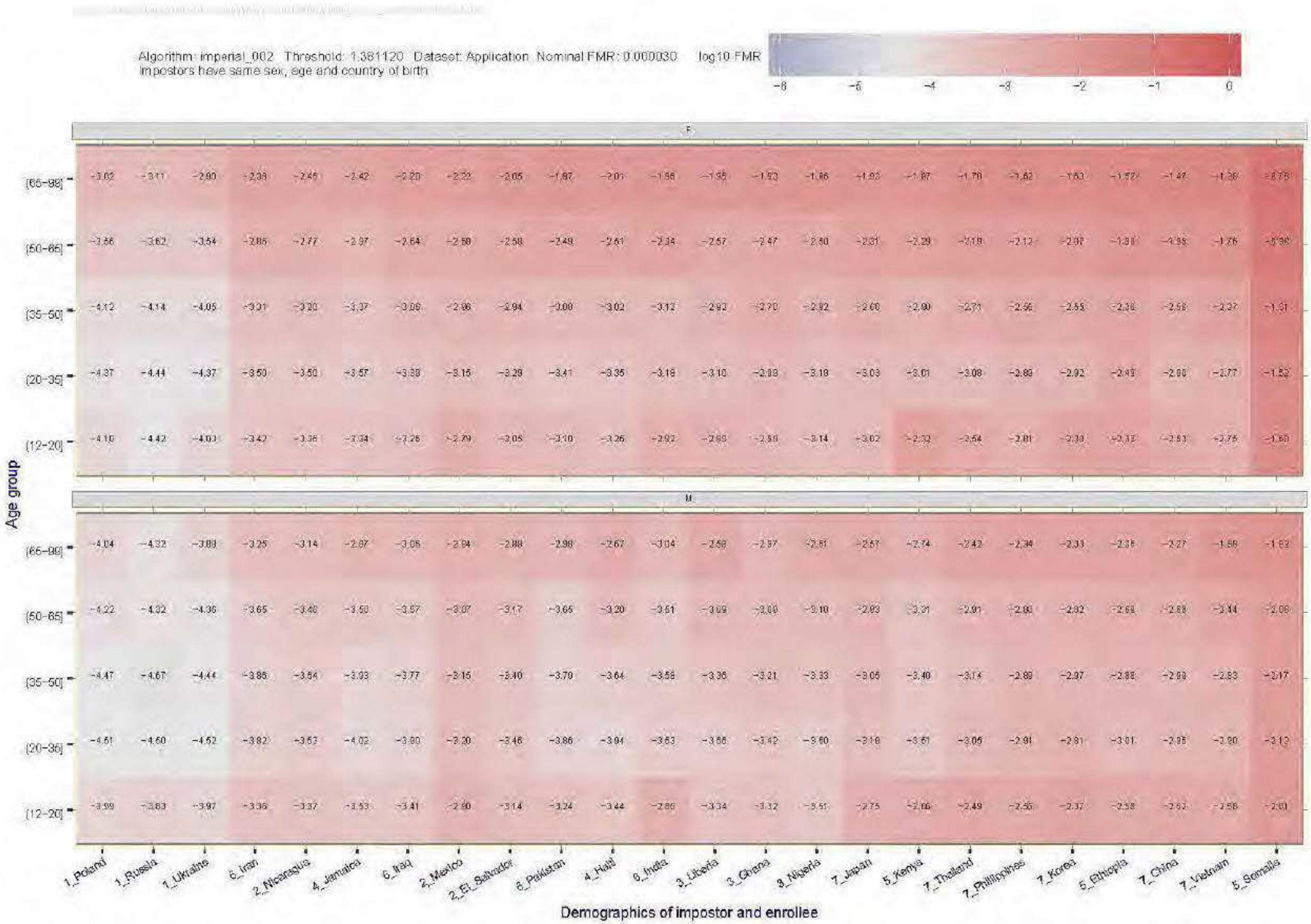


Figure 16: For application photos, the The heatmap shows one-to-one false match rates for same-sex, same-age and same-country of birth imposters, broken out by age and country. The text value is $\log_{10}$(FMR) with large negative values encoding superior false match rates. Each cell depicts FMR on a logarithmic scale. The text value is $\log_{10}$(FMR) with large negative values encoding superior false match rates. Annex 11 contains the corresponding figure for all algorithms.

5 False negative differentials in verification

5.1 Introduction

False negatives occur in biometric systems when samples from one individual yield a comparison score below a threshold. This will occur when the features extracted from two input photographs are insufficiently similar. Recall that face recognition is implemented as a differential operator: two samples are analyzed and compared. So a false negative occurs when two from the same face appear different to the algorithm.

5.2 Tests

This section gives empirical quantification of the variation in false negative rates across demographics. We base this on recognition results from three one-to-one verification tests:

- ▷ **Mugshot - Mugshot**: In the first test we look for demographic effects in the groups defined by the sex and race labels provided with these United States images - see Annex 1 .
- ▷ **Application - Application photo:** We consider also a high quality dataset collected from subjects hailing from twenty four countries in seven global regions.
- ▷ **Application - Border crossing photo:** As discussed in Annex 4 , the border crossing photos are collected under time constraints, in high volume immigration environments. The photos there present classic pose and illumination challenges to algorithms.

5.3 Metrics

The metrics appropriate to verification have been detailed in section 3.1. These are related to particular applications in Figure 2. The dicussion in subsequent sections centers on false non-match rates at particular thresholds, i.e. FNMR(T).

5.4 Results

Figure 17 summarizes the false non-match rates for the 52 most accurate algorithms comparing mugshot photos. It does this for each of four race categories and two sexes[13]. Figure 18 takes the same approach but for 20 countries of birth and two age groups (over/under 45). It summarizes comparison of high quality immigration

[13] See Annex 1 for descriptions of the images and metadata.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

Figure 17: For mugshot comparisons, the figure shows the distribution of FNMR values over the 52 most accurate verification algorithms, by sex and race. The threshold was set for each algorithm to achieve FMR = 0.00001 over all imposter comparisons. The line within each box is the median over those algorithms; the box itself spans the interquartile range (26 algorithms) and the lines here extend to minimum and maximum values. The small box on the left side indicates the accuracy for best algorithm overall, on this dataset alphaface-001.

application photos with lower quality border crossing photos. These are described in Annex 2 and Annex 4 respectively.

We make the following observations.

- ▷ FNMR is absolutely low: In one-to-one verification of mugshots, the best algorithms give FNMR below 0.5% at the reasonably stringent FMR criterion of 0.00001. FNMR is generally below 1% with exceptions discussed below. For the more difficult application-border crossing comparisons, the best algorithm almost always gives FNMR below 1%. These error rates are far better than the gender-classification error rates that spawned widespread coverage of bias in face recognition. In that study [5], two algorithms assigned the wrong gender to black females almost 35% of the time. The recognition error rates here, even from middling algorithms, are an order of magnitude lower. Thus, to the extent there are demographic differentials, they are much smaller than those that (correctly) motivated criticisms of the 2017-era gender classification algorithms.

- ▷ FNMR in African and African American subjects: In domestic mugshots, the lowest FNMR in images of subjects whose race is listed as black. However, when comparing high-quality appliction photos with border-crossing images, FNMR is often highest in African born subjects. We don't formally measure contrast or brightness in order to determine why this occurs, but inspection of the border quality images shows underexposure of dark skinned individuals often due to bright background lighting in the border crossing environment. In mugshots this does not occur. In neither case is the camera at fault.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

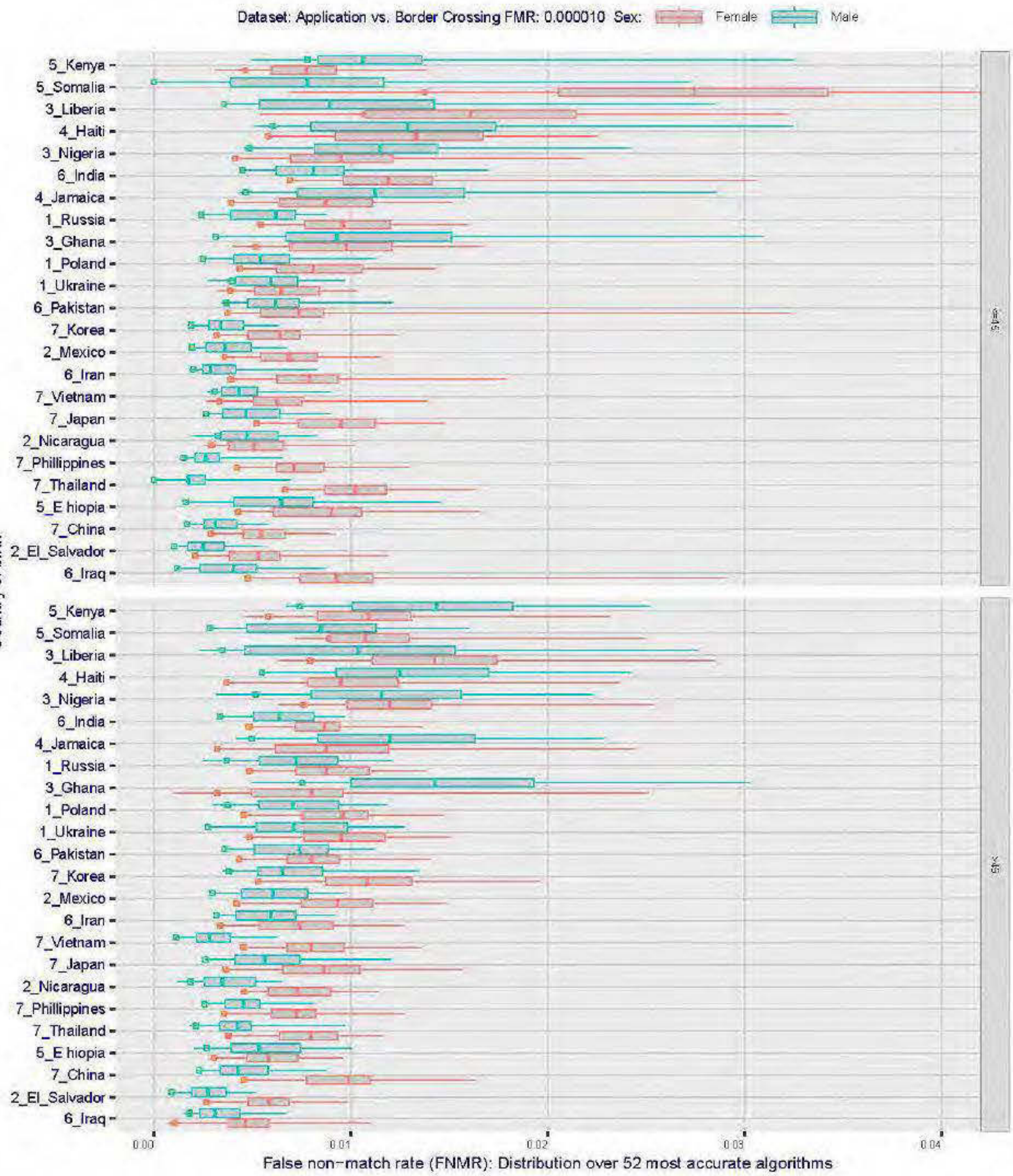


Figure 18: For the application - border crossing photo comparisons, the boxplots show the distribution of FNMR values over the 52 most accurate algorithms, by sex, country of birth, and age group. The threshold was set for each algorithm to achieve FMR = 0.00001 over all imposter comparisons. The line within each box is the median over those algorithms; the box itself spans the interquartile range (26 algorithms) and the lines here extend to minimum and maximum values. The small box on the left side indicates the accuracy for best algorithm overall, on this dataset visionlabs-007.

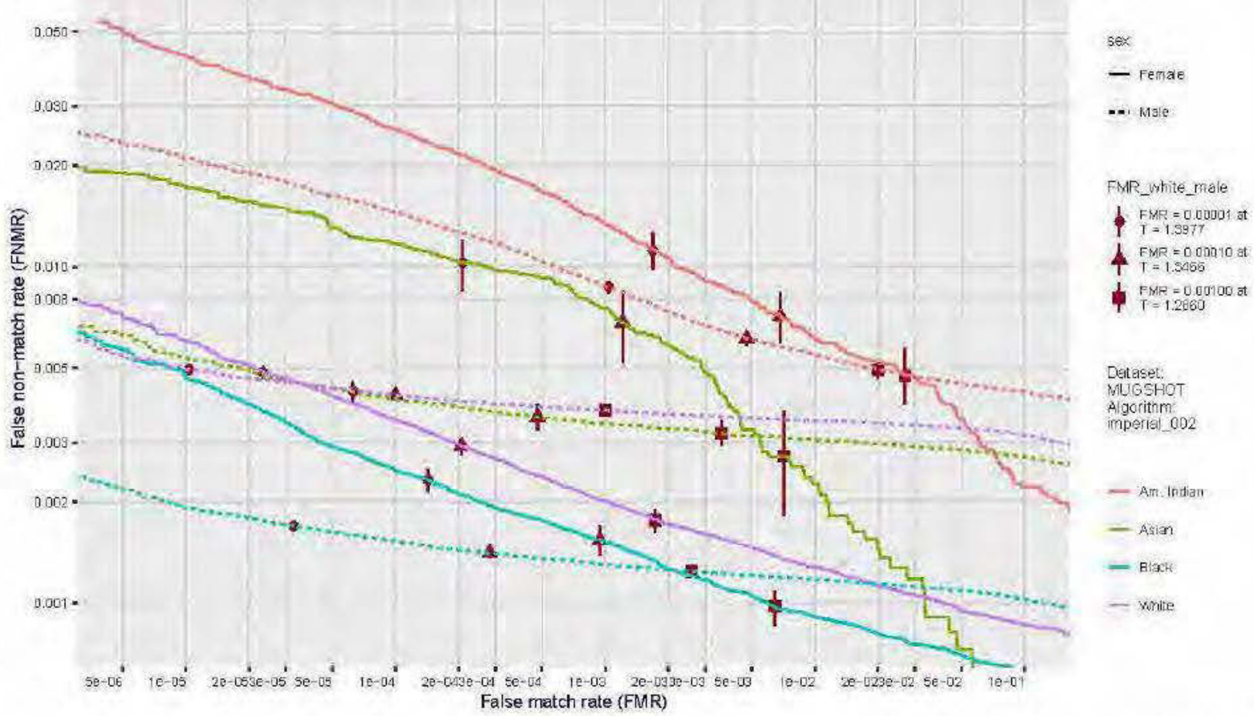


Figure 19: For one algorithm verifying mugshot images, the error tradeoff characteristics show false non-match vs. false match rates. The FMR estimates are computed for same-sex and same-race imposter comparisons. Each symbol (circle, triangle, square) corresponds to a fixed threshold - their vertical and horizontal displacements reveal, respectively, differences in FNMR and FMR between demographic groups. The vertical line through each symbol indicates uncertainty related to sample size - it spans 95% of bootstrap samples of the genuine scores. Annex 12 *contains the corresponding figure for all algorithms.*

▷ Women give higher FNMR: In most cases, algorithms give higher false non-match rates in women than men. Note that this is a marginal effect - perhaps 98% of women are still correctly verified - so the effect is confined to fewer than 2% of comparisons where algorithms fail to verify. It is possible that the error differences are due to relative prevalence some unknown covariate. There are some exceptions, however: In Kenya, Nigeria, Jamaica men give higher FNMR. This applies in Haiti and Ghana also but only for people aged 45 or over.

These aggregations of results over a large number of algorithms is intended to expose coarse differences between demographic groups. In so doing it hides that certain algorithms may differ from the trends evident in the Figure. Full error tradeoff characteristics appear in Annex 12 .

The false negative results for law enforcement images apply to high quality mugshots, collected with deliberate consideration of standards. When image quality degrades, false negatives are expected to increase. We next consider results for the comparison of high quality Annex 2 application reference photos with Annex 4 border crossing images collected in a less controlled environment under some (implicit) time constraint. We report

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

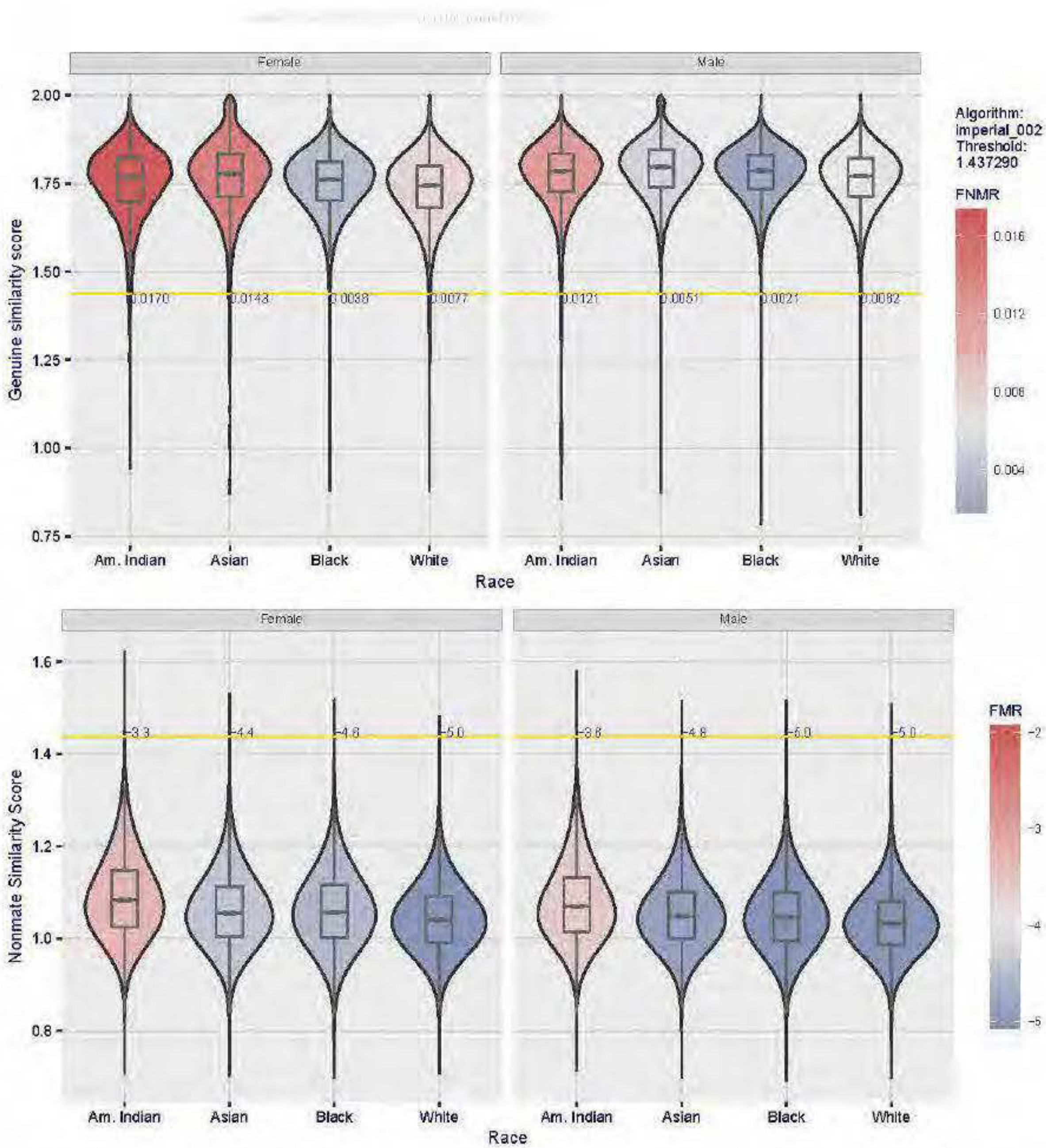


Figure 20: For one algorithm verifying mugshots, the violin plots show native similarity score distributions. The horizontal line shows the threshold that gives FMR = 0.0001 over all the imposter pairs. The imposters have the same sex and race. The upper figure shows genuine scores and the color indicates FNMR at the given threshold on a linear scale. The lower figure shows imposter scores with color indicating FMR on a logarithmic scale. FMR values below 10^{-5} are pinned to that value. Annex 15 contains the corresponding figure for all algorithms.

results in two ways:

▷ Per algorithm: Annex 14 shows FNMR by country of birth for two sexes and two age groups (above and below age 45).

▷ As Figure 22 heatmap showing results for all algorithms and all countries of birth. Each FNMR is the arithmetic mean of the four FNMR estimates for male and female and age over and under 45. The rows of the figure are sorted in order of mean FNMR, the mean being taken over all twenty four countries. The columns of the figure are sorted in order of mean FNMR from the 50 most accurate algorithms - this statistic was chosen so that high FNMR estimates from poor algorithms did not skew the results.

From these figure we note the following:

- **Wide variation across algorithms:** False non-match rates range from near 0.1% up to above 10%. This two-orders-of-magnitude range shows that some algorithms are intolerant of the quality problems inherent in the image the border crossing images. These problems are: low contrast, non-centered and cropped faces, non-frontal pose, and poor resolution, in part due to poor compression.

- **The most accurate algorithms give low FNMR:** The most accurate algorithms given FNMR below 1% for almost all countries and demographic groups. For example, the Visionlabs-007 algorithm has outliers only for Liberian and Somali women under the age of 45, for whom FNMR is below 1.4%.

- **Lower variation across countries:** For the more accurate algorithms, false non-match rates generally range by a factor of two or three from the left side of Figure 22 to the right i.e. FNMR in El Salvador is almost always lower than that in Somalia.

- **No clear patterns by age and sex:** By considering the Figures of Annex 14 , the differences between the over- and under-45s is often small, varies by country and by algorithm. However, broad statements do not mean that certain algorithms do not exhibit demographic differentials.

- **Higher FNMR in subjects from Africa and the Caribbean:** The heatmap is constructed with countries appearing in order of the mean FNMR over the fifty most accurate algorithms. This reveals higher FNMR in Africa and the Caribbean. After those two regions, the next highest FNMR is in the Eastern Europe countries.

The low error rates stem from efforts over the last decade to train algorithms that are invariant to nuisance variables such as non-frontal pose and poor contrast. The absolute magnitude of FNMR drives inconvenience. In many applications, any subject experiencing a false rejection could make a second attempt at recognition.

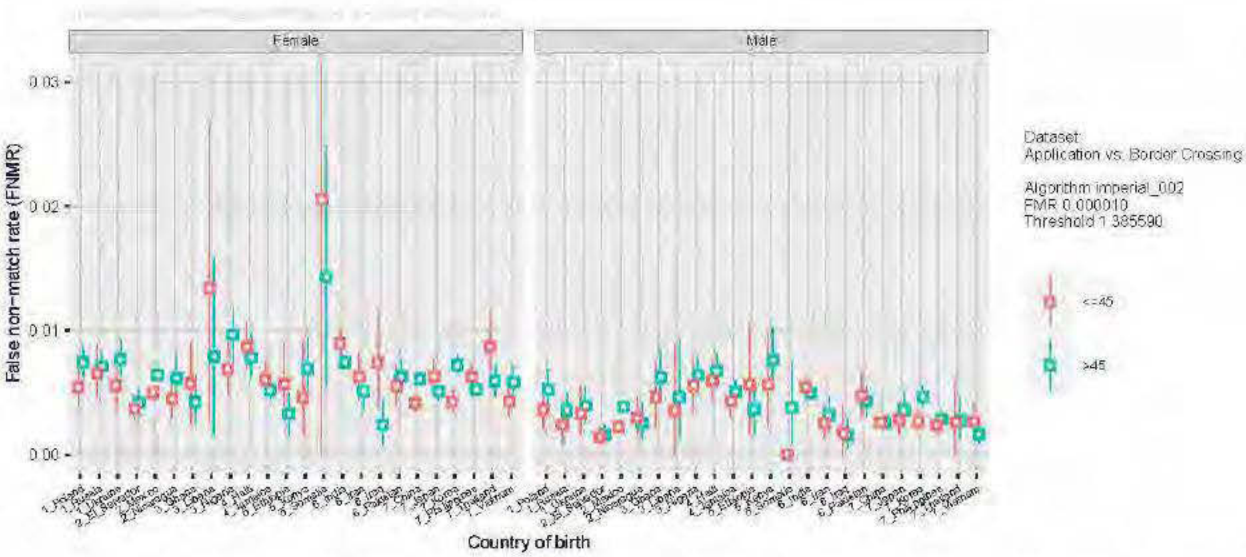


Figure 21: For 24 countries the figure shows false negative rates when the reference algorithm is used to compare two photos of subjects from the countries identified in the respective columns. The square box gives the median false non-match rate computed over 2000 bootstrap resamples of the genuine scores. The ends of the line span 95% of those re-samples, thereby giving a measure of uncertainty in the FNMR estimate. The threshold is set to a fixed value everywhere; it is the lowest value that gives FMR $\leq$ 0.00001. Annex 14 contains the corresponding figure for all algorithms.

Why these effects occur would require some multivariate analysis of image- and subject-specific properties. We suggest that analysis might start with measurement of image related quantities from the digital images to include such as contrast, intensity, areas of over and under exposure, presence of cropping, and head orientation. For tools, mixed-effects regression models could be an initial starting point [4] but such work would need to address correlation between quantities such race and contrast. We have not yet initiated such work and it is possible that such analysis would be incomplete due to influential but unknown covariates. In particular, given the border crossing images were collected with cameras mounted at fixed height and are steered by the immigration officer toward the face it is possible that subject height influences genuine matching scores. For example very tall subjects might be subject be underexposed because strong ceiling lights in the background might cause underexposure. Inspection of failure cases invariably leads to insight in such cases. We have not yet conducted that work.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

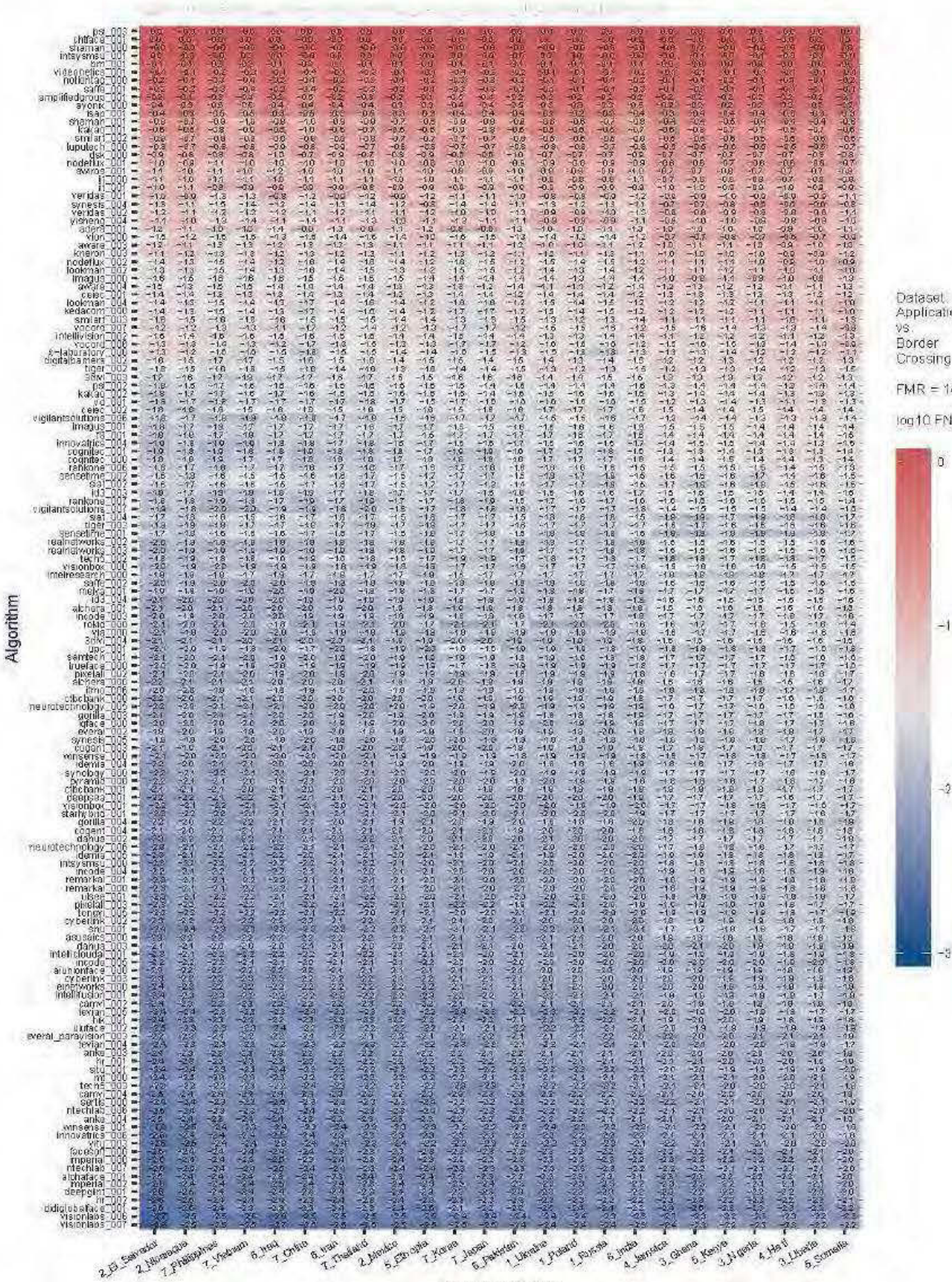


Figure 22: For 24 countries in seven regions the figure shows verification false non-match rates when the reference algorithm is used to compare two photos of subjects from the countries identified in the respective columns. The FNMR value is the mean over men/women and over/under age 45, so represents FNMR in situations where those four populations were balanced. The threshold is set to a fixed value everywhere; is is the lowest value that gives FMR $\leq$ 0.00001. Each cell depicts FNMR on a logarithmic scale. The text value is $\log_{10}$(FNMR) with large negative values encoding superior false match rates.

6 False negative differentials in identification

The three identification trials all use just mugshot photographs. They were conceived of to isolate specific demographic factors as follows.

- ▷ **Sex**: We construct a gallery containing 800 000 white men, and 800 000 white women, aged 20 - 40. We search that with mated probes taken in a different calendar year to the enrolled photo but no longer than 5 years after. We search with balanced sets of non-mate probes, also aged 20-40.
- ▷ **Sex**: We construct a gallery containing 500 000 black men, and 500 000 black women, aged 20 - 40. We search that with mated probes taken in a different calendar year to the enrolled photo but no longer than 5 years after. We search with balanced sets of non-mate probes, also aged 20-40.
- ▷ **Race**: We construct a gallery containing 800 000 black men, and 800 000 white men, aged 20 - 40. We search that with mated probes taken in a different calendar year to the enrolled photo but no longer than 5 years after. We search with balanced sets of non-mate probes, also aged 20-40.

More detail appears in Annex 16 . In each case the mated probes are used to measure false negative identification rate, and the nonmated probes are used to measure false positive identification rate. These tests all employ domestic mugshots, and only younger adults. Further work will extend analysis to a global population with more range in age.

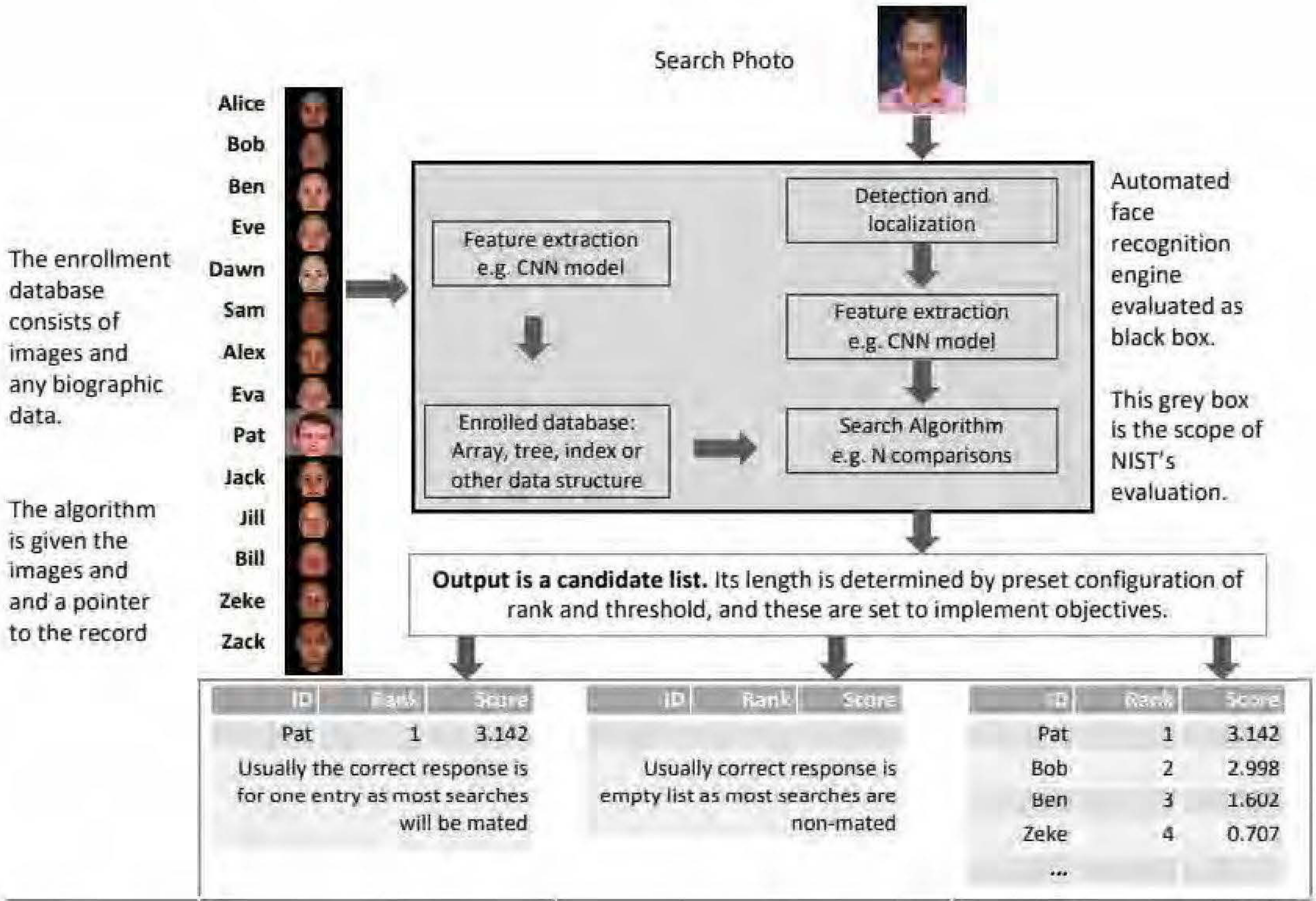
This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

6.1 Metrics

The metrics appropriate to identification have been detailed in section 3.2. These are related to particular applications in Figure 23 reflecting two modes of operation. The general metric FNIR(N, R, T) covers both as follows:

- ▷ **Investigation:** For investigators willing to traverse long candidate lists in pursuit of a lead, the metric FNIR(N, R, 0) is the proportion of missed mates when searching an N-enrollee gallery and considering the R most similar candidates without applying a threshold (T = 0). The utility of longer lists is shown by plotting FNIR vs. R.
- ▷ **Identification:** For those applications where a non-zero threshold is used to only return results when a search has a likely enrolled mate, the metric is FNMR(N, R, T). The use of thresholds $T > 0$ will suppress many false positives, but will also elevate false negatives, the tradeoff being shown as a plot of FNIR(T) vs. FPIR(T).

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280



	Positive identification	[illegible]e identification	Post-event investigation
Example	Access to a gym or cruise ship	Watchlist e.g. Detection of deportee or duplicate drivers license applications	Crime scene photos, or of detainee without ID documents.
Claim of identity	Implicit claim to be enrolled	Implicit claim to not be enrolled	No claim: Inquiry
Threshold	High, to implement security objective	High, to limit false positives	Zero
Num. candidates	1	0	L, set by request to algorithm
Human role	Review candidate to assist user in resolution of false negatives, or to detect impostor	Review candidate to determine false positive or correct hit	Review multiple candidates, refer possible hits to examiners see [26]
Intended human involvement frequency	Rare – approx. the false negative identification rate plus prior probability of impostor	Rare – approx. the false positive identification rate plus prior probability of an actual mate	Always
Performance metric of interest	**FNIR at low FPIR. See sec. 3.1, 3.2 and Tables 10, 19**	**FNIR at low FPIR. See sec. 3.1, 3.2 and Tables 10, 19**	**FNIR at ranks 1... 50, say. FPIR = 1. See sec. 3.2 and Table 12, 14, 16**

Figure 23: Identification applications and relevant metrics.

6.2 Results

Figures 24 and 25 show identification error rates for two algorithms. Plots for all algorithms are included in Annex 16 . In each case, the upper panels show FNIR vs R. The lower panels show FNIR vs. FPIR. We make the following observations

- Differentials by race in men: From the left-side panels, black men invariably give lower false negative identification rates than white men. This applies in the investigate and identification modes, and particularly for the more accurate algorithms. The differentials are often small, well below a factor of two. The are some exceptions including algorithms from 3DiVi, Aware, Eyedea, Idemia, Kedacom, Tevian and Vocord.
- Differentials between the sexes: Women invariably give higher false negative rates than men. This applies within both racial groups. There are exceptions, notably that searches of white women are more likely to produce the correct result in the top ranks than are search of men. This is less true for black women. A possible mechanism for this is available from section 4 verification results, namely that black women tend to produce high one-to-one false match rates. High non-mate scores may be displacing the correct black women from rank 1 position.
- Low FPIR is not attainable: The error tradeoff characteristics show a rapid increase FNIR as the threshold is increased to reduce FPIR. For example, in FNIR Figure 24, FNIR reaches 50% when FPIR is reduced to 0.0001. This is due to the presence of high scoring non-mates in the imposter searches. They can occur for several reasons. First, ground truth identity labeling errors in which photos of a person are in the database under multiple IDs. These cause apparent false positives. We discount this because the mugshot ground truth integrity is excellent, and underpinned by ten-print fingerprint matching. A second reason is the presence of twins in the population. Given the population represented by the dataset, we estimate a few percent of the United States adult population is present in the dataset. Given well documented twinning rates[14] [27], we expect twins to be in the data, both identical and, more commonly, fraternal. Siblings will be expected to give elevated similarities along the same lines.
- Higher false positive identification rates in black women: The lines connecting points of fixed threshold are often long and slanted in the error tradeoff plots in the center column of the bottom row - see Figure 24, for example. This is a common occurence revealing an order-of-magnitude increase in FPIR, with magnitudes varying by algorithm. Notably some algorithms do not exhibit this excursion. For example, the algorithm featured in Figure 25 gives much smaller excursions in FPIR.

[14]See the CDC's National Vital Statistics Report for 2017: https://www.cdc.gov/nchs/data/nvsr/nvsr67/nvsr67_08-508.pdf

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

Figure 24: For mugshot identification, the top row shows false negative identification "miss" rates as a function of rank, a metric appropriate to investigators traversing candidate lists for lead generation. The bottom row shows miss rates as a function of false positive identification rate, where a threshold is swept from a low value on the right to high values on the left. This metric is appropriate to organizations for which the volume of searches is high enough that they cannot afford labor to review results from every search. The left panels show the effect of race in young men. The center and right panels show difference between men and women, in black then white subjects respectively. The grey lines join points of equal threshold. The four thresholds are chosen to give FPIR of $\{0.0003, 0.003, 0.03, 0.3\}$ respectively for one baseline demographic, here white males. The figure applies to one algorithm, provided to NIST in August 2019. The corresponding figures for all identification algorithms appear in Annex 16 .

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

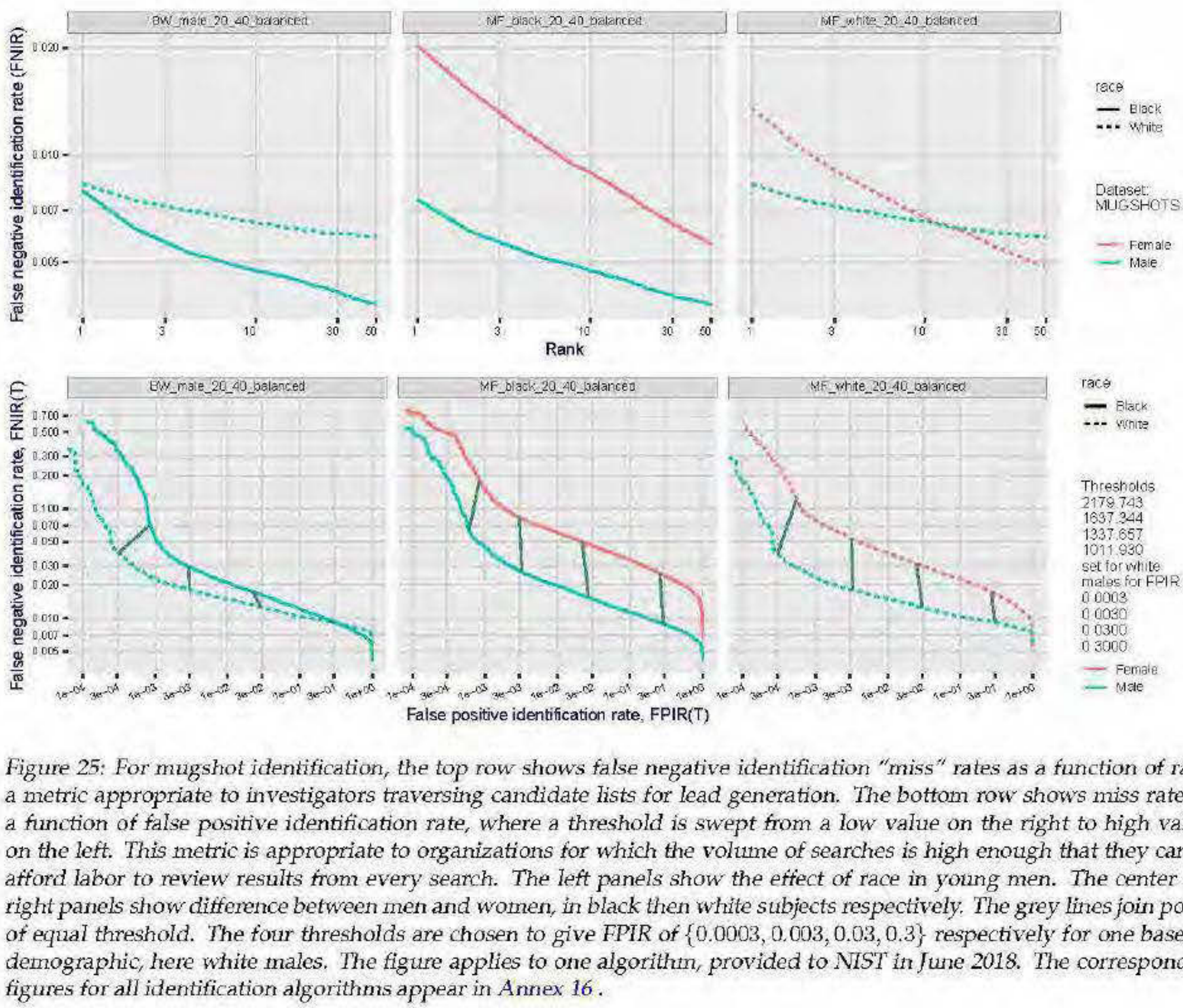


Figure 25: For mugshot identification, the top row shows false negative identification "miss" rates as a function of rank, a metric appropriate to investigators traversing candidate lists for lead generation. The bottom row shows miss rates as a function of false positive identification rate, where a threshold is swept from a low value on the right to high values on the left. This metric is appropriate to organizations for which the volume of searches is high enough that they cannot afford labor to review results from every search. The left panels show the effect of race in young men. The center and right panels show difference between men and women, in black then white subjects respectively. The grey lines join points of equal threshold. The four thresholds are chosen to give FPIR of $\{0.0003, 0.003, 0.03, 0.3\}$ *respectively for one baseline demographic, here white males. The figure applies to one algorithm, provided to NIST in June 2018. The corresponding figures for all identification algorithms appear in Annex 16 .*

7 False positive differentials in identification

The section addresses whether identification algorithms exhibit similar false positive differentials to verification algorithms. We first note that large-scale one-to-many identification deployments typically operate at false match rates much lower than those targeted in verification applications. It is typical in verification access control to target false match rates (FMR) between 0.00001 and 0.001, i.e. between one per hundred thousand and one per thousand. Identification applications, howveer, often enroll very large numbers of individuals numbering into the 10s or 100s of millions. If such systems are configured with thresholds aimed at producing false positive outcomes say one in 100 times, i.e. FPIR = 0.01, then the implied likelihood that a comparison will yield a false match is given by this formula

$$\text{FMR} = \frac{FPIR}{N} \tag{10}$$

where N is the size of the enrolled population. With FPIR= 0.01, and N= $10^6 s$ this formula implies FMR= 10^{-8}. The formula gives a first order equivalence of identification with verification: the former needs low false positive rates in large galleries. Metrics are discussed in section 3.

Some one-to-many search algorithms implement a 1:N search of a probe image as N 1:1 comparisons of the probe with the N enrolled items. This is followed by a sort operation which yields N candidates sorted in decreasing order of similarity. The result of that is returned in either of two ways: The system will return an operator-specified number of candidates, or it will return however many candidates are above an operator-specified threshold[15]. In the case where a threshold is used, the number of candidates returned will be a random-variable that is dependent on the image data itself.

Other algorithms do not implement 1:N search as N 1:1 comparisons. Instead they might employ a set of fast-search algorithms aimed at expediting search [2, 19, 21, 26]. These include various techniques to partition the enrollment data so that far fewer than N comparisons are actually executed. However, this does not mean that false positive occurences will be reduced because the algorithms are still tasked with finding the most similar enrollments.

For the three experiments listed in section 6, Figure 26 shows median scores returned by one identification algorithm when non-mated searches are conducted. It is clear that if a threshold is applied there will be demographic differences in the number of candidates returned, and in the score values. Such behavior applies to many algorithms - see Annex 17 .

This effect disappears in the algorithm featured in Figure 27. This is an important result because it implies much more equitable likelihoods of false positives. This is especially important result in negative identification

[15] The "operator-specified" parameters might sometimes be set by-policy, or by the manufacturer of the system.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

Figure 26: For searches of Asian, black, white men and women's faces into mixed galleries of mugshot photos the heatmaps show median similarity scores for candidates placed at rank 1 to 50. The upper four panels are produced in nonmated searches; the lower four from mated searches. The left-side panels are produced from searches into galleries with 12 000 000 people enrolled. The right-side uses galleries with N = 1 600 000 enrolled. The "lifetime consolidated" and "recent" labels refer to inclusion of multiple images per person, or just one - see [17].
Contrast the behavior here with that in Figure 27 and the corresponding figures for developers Aware, Idemia, NEC, Tevian, and Toshiba that are included in Annex 17 .

applications where the prior probability of a searched person actually being is in the database is low, e.g. card-sharp surveillance in a casino, or soccer hooligans at a sports game[16]. The lack of an effect on false positive identification rates is evident in Figure 25 where the grey lines join points of equal threshold. From left-to-right, the FPIR values for black and white males, black men and women, and white men and women are closely similar. The more normal behavior (see Figure 24 and Annex 16) is for larger shifts in false positive rates.

We now consider the implications for **investigative "lead generation" applications**. In such cases, algorithms return a fixed number of candidates and human reviewers compare the probe photo alongside each candidate gallery photo to determine if the photos are a match. In mugshot-mugshot searches the reviewer will very often look no further than rank 1 per the very high accuracy results documented in NIST Interagency Report

[16] For example, a recent news article noted the use of automated face recognition to search around 21 000 spectators at soccer games against a watch-list of about 50 people.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

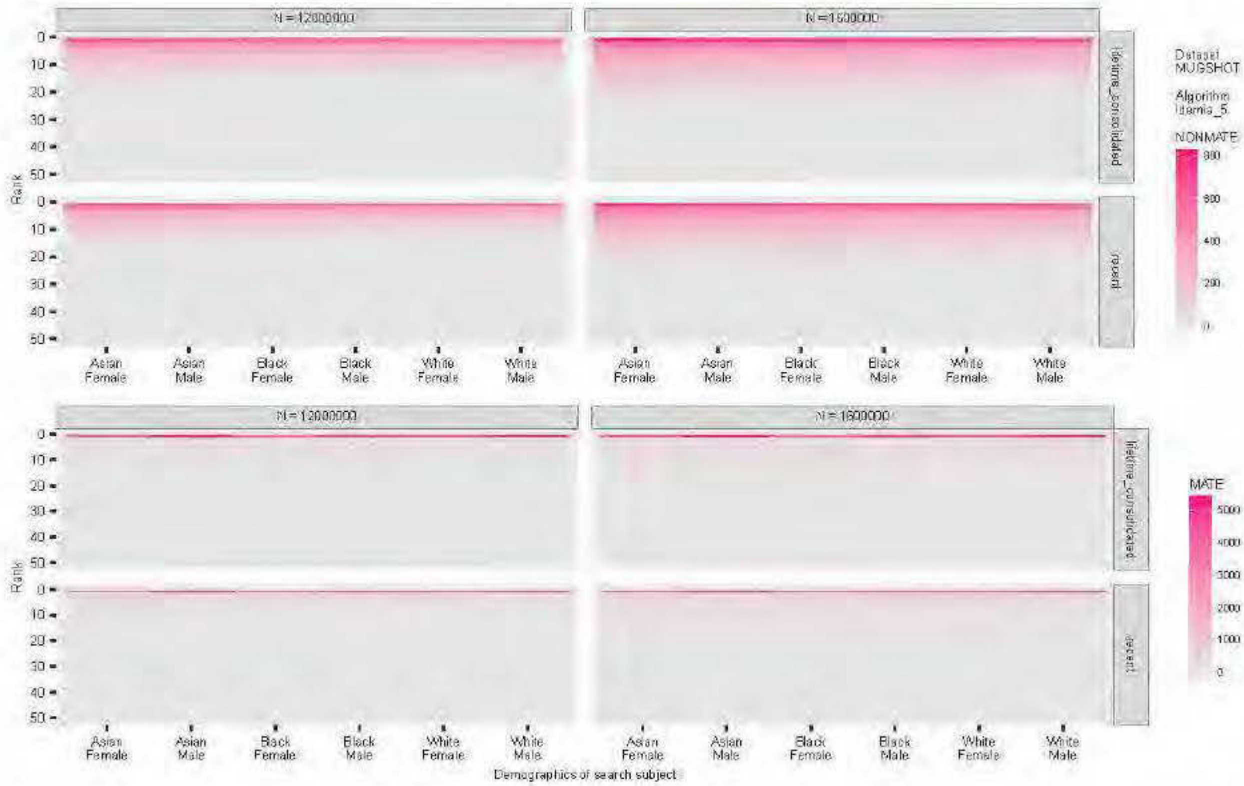


Figure 27: For searches of Asian, black, white men and women's faces into mixed galleries of mugshot photos the heatmaps show median similarity scores for candidates placed at rank 1 to 50. The upper four panels are produced in nonmated searches; the lower four from mated searches. The left-side panels are produced from searches into galleries with 12 000 000 people enrolled. The right-side uses galleries with N = 1 600 000 enrolled. The "lifetime consolidated" and "recent" labels refer to inclusion of multiple images per person, or just one - see [17]. The uniformity of the scores across demographic groups is in contrast to that evident in Figure 26 and many others in the Annex 17 compendium.

8271. That report also includes a workload measure summarizing the expected number of candidates that will have to be reviewed before a mate is located. A very important parameter in such applications, however, is the prior probability that a mate is actually present. In boarding a cruise ship for example, almost everyone attempting to board would be present in the gallery. In a casino application aimed at detecting "high rollers" the likelihood a patron of the casino is in that set is much lower. In such cases a human reviewer, if so employed, would in most searches review all say 50 candidates on the list. That's laborious and may not be tenable from an operations research perspective due to fatigue and reward factors in humans.

But in whatever circumstances human reviewers are tasked with reviewing candidate lists, how are demographic differentials such as those in Figure 26 expected to influence the human? The human will see fifty candidates regardless. However, if those candidates are accompanied by scores, presented as text in a GUI for example, the reviewer will see higher scores in the black female population and potentially elsewhere. Over time this may influence the human, though one earlier study [12] looked at cognitive bias issues in the human review of fingerprint search results, without demographic effects, and found scant evidence that scores influence the reviewer. That study did, however, find that just the order in which candidates are presented to reviewers affects both false positives and false negatives. For example, reviewers are more likely to miss (i.e. a false negative) a mated candidate that appears far down the candidate list. The issues involved in human review are beyond the scope of this document, but full consideration of *systems* comprised of automated face search algorithms and human reviewers is an experimental psychology, human factors and operations research issue.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

8 Research toward mitigation of false negatives

False negative error rates, and demographic differentials therein, are reduced in standards-compliant images. This motivates the following two research and development possibilities.

▷ **Improved standards compliance:** The ISO/IEC 19794-5 standard includes requirements regulating geometry and exposure. Recent research [24] noted that higher quality images, as determined by an automated quality assessment algorithm, yields a reduced false negative differential. While commercial packages exist for the automated assessment of quality, and NIST has an ongoing assessment of the underlying algorithms, rejection of single images on quality grounds can itself have demographic problems [1]. The ISO/IEC SC 37 biometrics subcommittee has recently initiated work on quality (ISO/IEC 29794-5 and 24357).

▷ **Face-aware cameras:** The same ISO/IEC committee has recently initiated work on specifications for capture subsystems that may require real-time face detection, pose estimation, and exposure measurement. Analogous "auto-capture" quality control mechanisms exist in iris and fingerprint scanners. That standard, ISO/IEC 24358, will be developed through 2020 with completion expected in 2021. Participation is open via national standardization groups.

Along similar lines further research into automated image quality assessment, and particularly specifications for closed-loop face-aware capture would prove valuable in averting low-contrast and over- and under-exposed images. Many enrollment operations still rely on documentary photography standards with cameras that are not detecting and metering off faces.

This work would be supported by research into two further topics:

Analysis: There is a need for improved models of demographic effects, particularly to how subject-specific properties including phenotypes, imaging artefacts and algorithms interact. Such models would extend work [9] in separating the relative contributions of at least, sex, age, race and height. Efforts to automatically estimate phenotypic information from images will involve algorithms that may themselves exhibit demographic differentials. Such work will need to address this possibility.

Information theoretic analysis: Given the potential for poorly illuminated photographs to produce false negatives, via under- or over-exposure of dark or light skin, an information theoretic approach to characterize algorithmic response to poor lighting would be useful for future standardization. In particular, the ISO/IEC 19794-5 standard has, since 2004, required portrait photos to have at least 7 bits of content in each color channel. Such work should quantify both false negative and false positive dependence.

9 Research toward mitigation of false positives

9.1 Summary

The threshold manipulation strategies described above would be irrelevant if the algorithm developer provided software with homogeneous false match rates. That will prove impossible as there will always be some distribution around a mean - the goal should be much more homogeneous false match rates than is currently the case.

9.2 Algorithm training

A longer-term mitigation is prompted by our observation that many algorithms developed in China do not give the elevated false positive rates on Chinese faces that algorithms developed elsewhere do. This affirms a prior finding of an "other-race effect" for algorithms [33] though that paper did not separate false positive from false negative shifts. This suggests that training data, or perhaps some other factor intrinsic to the development, can be effective at reducing particular false positive differentials. Thus, the longer-term mitigation would be for developers to investigate the utility of more diverse, globally derived, training data. Absent such data, developers might consider whether their cost functions can be altered to reduce differentials. One developer advanced such a concept in November 2018 [15].

9.3 Greater discriminative power

Face recognition algorithms measure similarity between face images. Facial appearance is partially determined by genes, the phenotypic expression of which determines skin tone and a large set of characteristics related to shape of the face. In NIST recognition tests [17], identical twins invariably cause false positives at all practical operational thresholds. Twins are characterized by very similar features given identical genes. Similarities in faces in fraternal twins [17] are expected to extend also to siblings (which also share half of the genes), and then to more distant relatives. In 2004, an algorithm was patented that can correctly distinguish twins [US Patent: US7369685B2]; it operates by extracting features from skin texture (adjacent to the nose, and above the eyebrows). This algorithm requires high resolution and, moreover, knowledge that any given image has that resolution. However, contemporary deployments of face recognition are very often based on processing of images at or below VGA spatial sampling rates (i.e., 480 x 640 pixel images), and this is often insufficient for skin texture to be viable. The human reviewer community has long specified much higher resolution for forensic purposes (see ANSI/NIST Face Acquisition Profiles).

9.4 Collection and use of face and iris

The texture of the human iris is known to have a structure that when imaged and processed by published feature extraction algorithms [11, 29] will correctly discriminate between identical twins [40] - something that contemporary marketplace face algorithms do not [17]. The reason for this appears to be that the iris features detected by automated algorithms are not genetically determined. However genetics research [25] does show iris textures have some genetic linkage, so a better characterisation of the tails of the impostor distribution is needed, at least for large scale one-to-many identification. Nevertheless, a 2019 DHS Science and Technology study noted that false positives are no higher within individuals of the same sex, age and race as they are across those groups [39]. As shown in Figure 4 and Annex 8 that is not the case for face recognition. NIST has near-term plans to investigate the impostor distribution in twins more fully.

Given the marketplace presence of multiple cameras that collect face and iris essentially simultaneously, one approach to consider for mitigation of false positive differentials in face recognition would be for face records to include adjunct iris images. The standards infrastructure is in place for this (ANSI/NIST Type 17, ISO/IEC 39794-6, and ICAO 9303 Data Group 4). This would afford very low false positive rates.

The apparent lack of genetic influence, and demonstrated low false match likelihoods, has been the primary property in establishing the use of the iris for the identification of individuals in large populations - most notably in the Indian National ID program Aadhaar. The iris recognition industry has multiple camera developers, multiple algorithm suppliers, and image interchange and quality standards that support interoperable recognition across cameras.

These aspects afford solutions to higher and heterogeneous false positive rates in face recognition. The first is simply to replace face with iris. There would be advantages and disadvantages to this - detailing and weighing those is beyond our scope here. However a second solution would be to augment face with iris, to produce a compound biometric "face-and-iris"[17]. This is made possible by the marketplace availability for at least a decade now of cameras that collect iris and face images essentially simultaneously. Recognition of the combined biometric would involve a particular kind of biometric fusion that in which both the face and iris must match (against respective thresholds) so as to limit false positives. This differs from some convenience-driven implementations that authenticate a person with either face or iris alone.

Use of iris in some applications, for example surveillance, is limited by the difficulty and expense of imaging the iris at long distances.

We don't mention fingerprints in this context because even though genetic influence is considered to be absent

[17] Such a compound biometric would conventionally still require collection of two images: First an iris image with near infrared illumination and the face image either entirely in ambient light, or ambient light with a near infrared component. The recognition of irises in purely visible-light images is highly problematic in brown-eyed people as melanin in the iris absorbs incident light at visible wavelengths.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

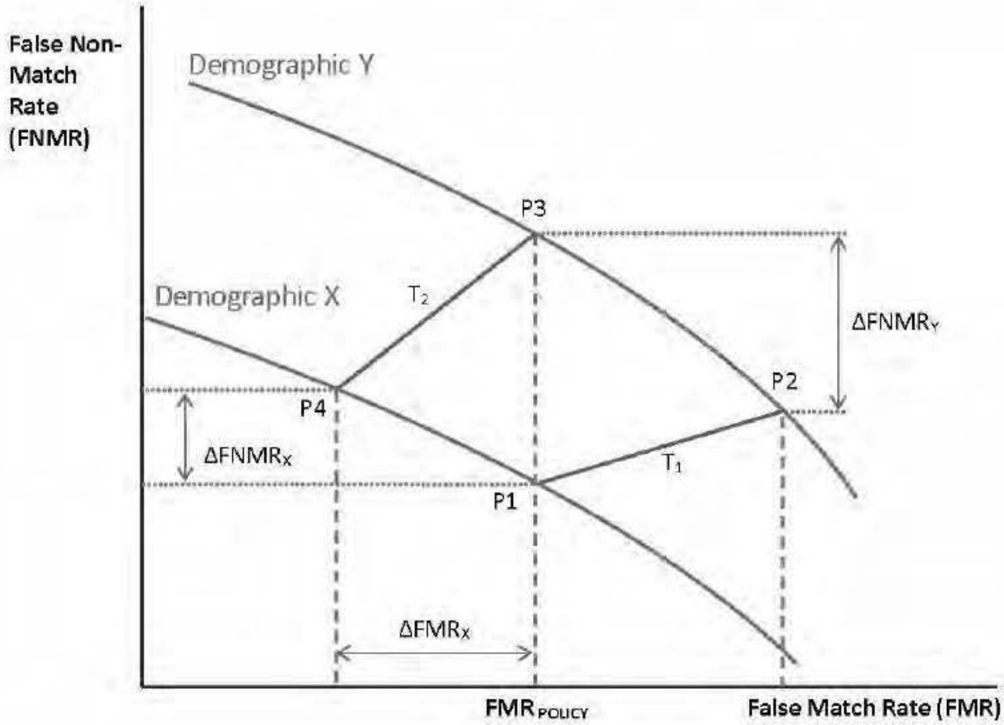


Figure 28: The figure shows the increases in FNMR implied by increasing the operating threshold to achieve the target FMR on the high-FMR demographic, Y.

or minimal, the collection of both fingerprint and face is not simultaneous.

9.5 Threshold elevation

We detail one mitigation of heterogeneous variable false match rates, and its consequences, as follows. The explanation uses a graphical construct based on the error tradeoff characteristics shown in Figure 28.

▷ We start with a target false match rate $\mathrm{FMR}_{\mathrm{POLICY}}$ that has been set to implement some security objective. This value, in a verification application might reasonably be set to say 1 in 5000 (i.e. 0.0002). This is implemented by setting a threshold T_1. Suppose that this threshold was perfectly calibrated for Demographic X i.e. $\mathrm{FMR}(T_1) = \mathrm{FMR}_{\mathrm{POLICY}}$. This corresponds to the point P1.

▷ Now suppose that we later discover, perhaps as a result of some biometric performance test or audit that, for some new group Demographic Y, that the observed false match rate at the fixed threshold T_1 is much higher, a factor of five say (0.001). This point P2 therefore represents therefore a failure to meet the original security objective for that group.

▷ To bring the overall system into policy compliance, the system owner consults the error tradeoff charac-

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

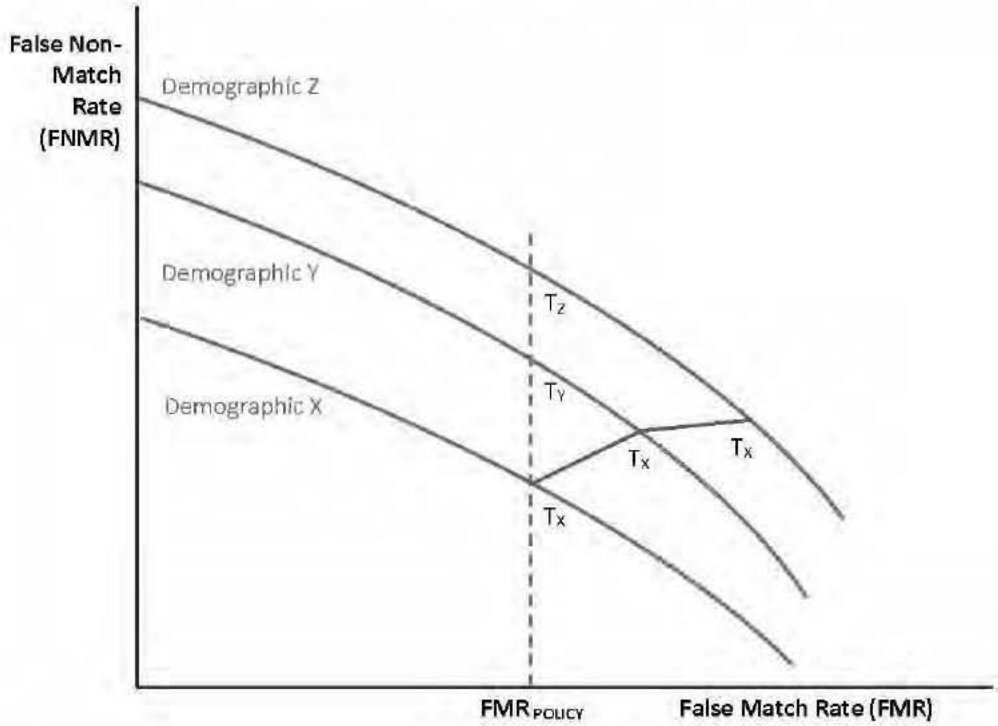


Figure 29: The figure shows the effect of setting thresholds to achieve the target FMR on demographics X and Y.

teristic for Demographic Y and notes that by elevating the threshold to T_2, the false match rate would be returned to policy compliance, at point P3.

▷ The effect of this however is that FNMR is necessarily elevated both demographic groups. This is because the new threshold T_2 is higher than T_1, and applies to all transactions from all demographics. These increases are shown as $\Delta FNMR_X$ and $\Delta FNMR_Y$ would have a magnitude that depends on the gradients of the error-tradeoff characteristics (which may differ). The only gain is a reduction in FMR for Demographic X, to a value which beats the original target policy.

Using this kind of construct, we see the benefit in having a biometric algorithm for which false match rates are homogenous i.e. do not vary (much) over any demographics.

The above argument assumes that the original high $FMR_Y(T_1)$ is indeed problematic. It may be tolerable in cases where individuals in that Demographic are rare, e.g. elderly persons entering a gym or nightclub. Any decision to not elevate the threshold to T_2 should be deliberated in the security context defined by threat, risk and cost.

9.6 Explicit thresholds for each demographic

In this section we discuss the suggestion [23] to address heterogeneous false match rates by assigning a threshold to each demographic. The proposal is for a verification system to set the threshold each time a subject executes a verification transaction tailoring it on the basis of who is using the system. Referencing Figure 29, this would correspond to adopting thresholds T_1 and T_2 (i.e. points P1 and P3) on-the-fly. How to do this presents a problem. Naively one could encode in an identity document (e.g. a passport) some indication of the demographic group (e.g. female, middle aged, south Asian) and the system would read this information, consult a lookup table, and set T accordingly. This would be effective for genuine legitimate users of the system. The security consequences of this are, however, more complicated. Consider what an imposter would do given knowledge that thresholds are variable.

- ▷ If the imposter were from a demographic for which the threshold is low, he would procure / steal a credential from somebody of the same age, sex and ethnicity. This would be typical behavior for any imposter. However, if particular countries passports were known to be used with low-thresholds, we'd expect genesis of a black-market for stolen credentials in those places.

- ▷ If the imposter were from a demographic for which the threshold is high he might procure / steal a credential from somebody in one of the low-threshold demographics, matching age and sex minimally the same sex. To better induce a false match the imposter would still need to have the same age, sex and ethnicity. This would be typical behavior anyway.

  Note that societal construction will often naturally afford opportunities for imposters to have access to identity credentials from other persons who, naturally, have the same ethnicity, sex and age group.

Another aspect to this approach is that it shifts responsibility for threshold management to the system owner rather than the developer. That may sound fully appropriate but imposes two responsibilities on the operator: First, figuring out what the thresholds should be via some appropriate testing, and secondly to implement the strategy with capture of demographic information and use of that in software.

References

[1] December 2016. https://www.telegraph.co.uk/technology/2016/12/07/robot-passport-checker-rejects-asian-mans-photo-having-eyes/.

[2] Artem Babenko and Victor Lempitsky. Efficient indexing of billion-scale datasets of deep descriptors. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[3] L. Best-Rowden and A. K. Jain. Longitudinal study of automatic face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(1):148–162, Jan 2018.

[4] J. Ross Beveridge, Geof H. Givens, P. Jonathon Phillips, and Bruce A. Draper. Factors that influence algorithm performance in the face recognition grand challenge. *Computer Vision and Image Understanding*, 113(6):750–762, 2009.

[5] Joy Buolamwini. Gender shades: Intersectional phenotypic and demographic evaluation of face datasets and gender classifiers. Technical report, MIT Media Lab, 01 2017.

[6] J. Campbell and M. Savastano. Iso/iec 22116 identifying and mitigating the differential impact of demographic factors in biometric systems. Technical report, ISO/IEC JTC 1, SC 37, Working Group 6, http://iso.org/standard/72604.html, 11 2018.

[7] Jacqueline G. Cavazos, Eilidh Noyes, and Alice J. O'Toole. Learning context and the other-race effect: Strategies for improving face recognition. *Vision Research*, 157:169 – 183, 2019. Face perception: Experience, models and neural mechanisms.

[8] Jacqueline G. Cavazos, P. Jonathon Phillips, Carlos D. Castillo, and Alice J. O'Toole. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? In *https://arxiv.org/abs/1912.07398*, 12 2019.

[9] Cynthia Cook, John Howard, Yevgeniy Sirotin, Jerry Tipton, and Arun Vemury. Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, PP:1–1, 02 2019.

[10] White D., Kemp R. I., Jenkins R., Matheson M, and Burton A. M. Passport officers errors in face matching. *PLoS ONE*, 9(8), 2014. e103510. doi:10.1371/journal. pone.0103510.

[11] J. Daugman. How iris recognition works. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):21–30, Jan 2004.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8280

[12] Itiel Dror and Kasey Wertheim. Quantified assessment of afis contextual information on accuracy and reliability of subsequent examiner conclusions. Technical Report 235288, National Institute of Justice, July 2011.

[13] H El Khiyari and Wechsler H. Face verification subject to varying (age, ethnicity, and gender) demographics using deep learning. *Journal of Biometrics and Biostatistics*, 7:323, 11 2016. doi:10.4172/2155-6180.1000323.

[14] C. Garvie, A. Bedoya, and J. Frankle. The perpetual line-up: Unregulated police face recognition in america. Technical report, Georgetown University Law School, Washington, DC, 10 2018.

[15] Stéphane Gentric. Face recognition evaluation @ idemia. In *Proc. International Face Performance Conference, National Institute of Standards and Technology NIST, Gaithersburg, MD*, November 2018.

[16] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test (frvt) part 1: Verification. Interagency Report DRAFT, National Institute of Standards and Technology, October 2019. https://nist.gov/programs-projects/frvt-11-verification.

[17] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test (frvt) part 2: Identification. Interagency Report 8271, National Institute of Standards and Technology, September 2019. https://doi.org/10.6028/NIST.IR.8271.

[18] Patrick Grother, George W. Quinn, and Mei Ngan. Face recognition vendor test - still face image and video concept, evaluation plan and api. Technical report, National Institute of Standards and Technology, 7 2013. http://biometrics.nist.gov/cs_links/face/frvt/frvt2012/NIST_FRVT2012_api_Aug15.pdf.

[19] Feng Hao, John Daugman, and Piotr Zielinski. A fast search algorithm for a large fuzzy database. *IEEE Transactions on Information Forensics and Security*, 3(2):203–212, 2008.

[20] John J. Howard, Yevgeniy Sirotin, and Arun Vemury. The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance. In *Proc. 10-th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2019, Tampa Florida,USA*, September 2019.

[21] Masato Ishii, Hitoshi Imaoka, and Atsushi Sato. Fast k-nearest neighbor search for face identification using bounds of residual score. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 194–199, Los Alamitos, CA, USA, May 2017. IEEE Computer Society.

[22] B. F. Klare, Burge M. J., Klontz J. C., Vorder Bruegge R. W., and Jain A. K. Face recognition performance: Role of demographic information. *IEEE Trans. on Information Forensics and Security*, 7(6):1789–1801, 9 2012.

[23] K. S. Krishnapriya, Kushal Vangara, Michael C. King, Vitor Albiero, and Kevin Bowyer. Characterizing the variability in face recognition accuracy relative to race. *CoRR*, abs/1904.07325, 2019. http://arxiv.org/abs/1904.07325.

[24] K. S. Krishnapriya, Kushal Vangara, Michael C. King, Vitor Albiero, and Kevin Bowyer. Us study: better image quality could cut face system bias. *Biometric Technology Today*, 2019(5):11 – 12, 2019.

[25] Mats Larsson, David L. Duffy, Gu Zhu, Jimmy Z. Liu, Stuart Macgregor, Allan F. McRae, Margaret J. Wright, Richard A. Sturm, David A. Mackey, Grant W. Montgomery, Nicholas G. Martin, and Sarah E. Medland. GWAS findings for human iris patterns: Associations with variants in genes that influence normal neuronal pattern development. *American Journal of Human Genetics*, 89(2):334–343, August 2011.

[26] Yury A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *CoRR*, abs/1603.09320, 2016.

[27] Joyce A. Martin, Brady E. Hamilton, Michelle J.K. Osterman, Anne K. Driscoll, , and Patrick Drake. National vital statistics reports. Technical Report 8, Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System, Division of Vital Statistics, November 2018.

[28] Dana Michalski, Sau Yee Yiu, and Chris Malec. The impact of age and threshold variation on facial recognition algorithm performance using images of children. In *International Conference on Biometrics*, February 2018.

[29] D. M. Monro, S. Rakshit, and D. Zhang. Dct-based iris recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):586–595, April 2007.

[30] Vidya Muthukumar. Color-theoretic experiments to understand unequal gender classification accuracy from face images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[31] Vidya Muthukumar, Tejaswini Pedapati, Nalini Ratha, Prasanna Sattigeri, Chai-Wah Wu, Brian Kingsbury, Abhishek Kumar, Samuel Thomas, Aleksandra Mojsilovic, and Kush R. Varshney. Understanding unequal gender classification accuracy from face images. *CoRR*, abs/1812.00099, November 2018.

[32] P. Jonathon Phillips, J. Beveridge, Bruce Draper, Geof Givens, Alice O'Toole, David Bolme, Joseph Dunlop, Yui Lui, Hassan Sahibzada, and Samuel Weimer. The good, the bad, and the ugly face challenge problem. *Image and Vision Computing*, 30:177185, 03 2012.

[33] P. Jonathon Phillips, Fang Jiang, Abhijit Narvekar, Julianne Ayyad, and Alice J. O'Toole. An other-race effect for face recognition algorithms. *ACM Trans. Appl. Percept.*, 8(2):14:1–14:11, February 2011.

[34] P. Jonathon Phillips, Amy N. Yates, Ying Hu, Carina A. Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G. Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, Jun-Cheng Chen, Carlos D. Castillo, Rama Chellappa, David White, and Alice J. O'Toole. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018.

[35] P.J. Phillips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone. Face recognition vendor test 2002. Evaluation Report IR 6965, National Institute of Standards and Technology, www.itl.nist.gov/iad/894.03/face/face.html or www.frvt.org, March 2003.

[36] Inioluwa Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Conference on AI, Ethics and Society*, pages 429–435, 01 2019.

[37] K. Ricanek and T. Tesafaye. Morph: a longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 341–345, April 2006.

[38] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. International Conference on Learning Representations*, volume https://arxiv.org/abs/1409.1556v6, 2015.

[39] Yevgeniy Sirotin. A comparison of demographic effects in face and iris recognition. Technical report, Iris Experts Group, Gaithersburg, MD, 6 2019.

[40] Zhenan Sun, Alessandra Paulino, Jianjiang Feng, Zhenhua Chai, Tieniu Tan, and Anil Jain. A study of multibiometric traits in identical twins. *Proc. of the International Society of Optical Engineering*, 7667, 04 2010.

[41] Darrell M. West. 10 actions that will protect people from facial recognition software. Technical report, Brookings Institution, Artificial Intelligence and Emerging Technology Initiative, Washington, DC, 10 2019.

[42] David White, James D. Dunn, Alexandra C. Schmid, and Richard I. Kemp. Error rates in users of automatic face recognition software. *PLoS ONE*, October 2015.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

NISTIR 8381

Face Recognition Vendor Test (FRVT)

Part 7: Identification for Paperless Travel and Immigration

Patrick Grother
Austin Hom
Mei Ngan
Kayee Hanaoka
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8381

National Institute of
Standards and Technology
U.S. Department of Commerce

NISTIR 8381

Face Recognition Vendor Test (FRVT)

Part 7: Identification for Paperless Travel and Immigration

Patrick Grother
Austin Hom
Mei Ngan
Kayee Hanaoka
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8381

July 2021

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
James K. Olthoff, Performing the Non-Exclusive Functions and Duties of the Under Secretary of Commerce for Standards and Technology & Director, National Institute of Standards and Technology

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

DISCLAIMER

Specific hardware and software products identified in this report were used in order to perform the evaluations described in this document. In no case does identification of any commercial product, trade name, or vendor, imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose.

INSTITUTIONAL REVIEW BOARD

The National Institute of Standards and Technology's Research Protections Office reviewed the protocol for this project and determined it is not human subjects research as defined in Department of Commerce Regulations, 15 CFR 27, also known as the Common Rule for the Protection of Human Subjects (45 CFR 46, Subpart A).

ACKNOWLEDGMENTS

The authors are grateful for the support and collaboration of the U.S. Customs and Border Protection (CBP) component of the Department of Homeland Security.

Additionally we are indebted to staff at DHS' Science & Technology Directorate (S&T) and Office of Biometric Identity Management (OBIM) for discussions and image data that supports this work.

RELEASE NOTES

This report will be updated periodically with results for new algorithms, new analyses, and new datasets as they become available.

The report is open for comment - correspondence should be directed to frvt@nist.gov.

Executive Summary

We investigate the use of one-to-many facial recognition in airport transit settings in which travelers' faces are matched against galleries of individuals expected to be present. We primarily consider the case where face recognition serves double-duty for access control (to an aircraft) and facilitation (of recording a visa-holder's departure from a country). This is done in a paperless mode in which a boarding pass (something you have) is replaced with presentation of a biometric (something you are) to a camera, representing an implicit claim to be entitled to board. We describe how such systems can fail, discussing errors during gallery creation, photo capture at boarding, attack detection, and face matching. We discuss how errors might be estimated, citing relevant standards, and their consequences.

We quantify face matching errors by simulating departing flights, populating galleries with an airport ENTRY photo of 420 travelers, then measuring accuracy by running searches of EXIT photos. We repeat this with galleries populated with multiple photos per person, and with galleries as large as 42000, modelling the same concept of operations but at a centralized airport checkpoint. We report that accuracy varies greatly across algorithms, that use of multiple images per person reduces errors considerably, and that error rates when searching 42000-person galleries are often three times higher than in 420-person galleries, but still sometimes below 1%. We consider demographics, and note that for the more accurate algorithms, error rates are so low that accuracy variations across sex and race are insignificant. We include additionally a discussion of how our accuracy estimates might differ from those measured operationally due to by factors that we could not control, such as camera type and imaging environment.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

Technical Summary

Background: One-to-many biometric search systems are discussed in their role of positive and negative identification - the former refers to the expectation that person in a probe sample is present in the database (as in access to an office) while the latter presumes the person is not (as in compulsive gamblers entering a casino). The distinction is useful because the applications differ in their tolerance for false negatives and false positives. This report addresses the positive use of one-to-many facial recognition in airport transit settings in which travelers' faces are matched against galleries of individuals expected to be present. We primarily consider the case where face recognition serves double-duty for access control (to an aircraft) and facilitation (of recording a visa-holder's exit).

In late 2018 the United States commenced a pilot of face-based confirmation of departure system in which passengers boarding an aircraft make cooperative presentations to a camera and the captured photos are immediately searched against a gallery comprised of photos of persons expected on the flight. This process is intended to biometrically bind the traveler to the departure. A positive biometric match in used two ways: First, by the airline, to grant access to the aircraft in lieu of a boarding-pass presentation; second, by passport control authorities to record the departure from the United States of in-scope passengers (e. g. visa holders), notionally replacing the long-standing airline manifest-based biographic process.

Overview: This report summarizes three NIST activities: First, to describe the biometric aspects of the traveler departure application and factors that are expected to affect its performance; second, to document results from running offline simulations in which recent accurate face recognition algorithms are applied to actual ENTRY and EXIT images with the goals of establishing a methodology, estimating accuracy, and exposing some factors that will affect those estimates; third to consider the use of face recognition at other airport touchpoints where higher populations are expected.

EXIT simulations: We simulate traveler EXIT by preparing 567 galleries each containing exactly 420 individuals representing the population expected on a flight. The individuals are not selected by age or sex. They are selected to have the same region of travel document (for example, South America or East Asia). We search each gallery with a fixed set of actual 132 931 EXIT photos using recent commercial one-to-many face recognition search engines. Notably we do not have camera, location and timestamp metadata so we cannot "replay" biometric boarding of actual flights. We

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

include this and other caveats in section 5.

Algorithms: Our EXIT simulations make use of one-to-many search algorithms submitted to NIST's ongoing Face Recognition Vendor Test between mid 2018 and April 2021. These algorithms are prototypes from the R&D laboratories of commercial developers of face recognition. These include two variants from the incumbent provider to the face matching facility used in the U.S., including the NEC-3 algorithm that was broadly the most accurate algorithm evaluated in 2018 as reported in NIST Interagency Report 8271 [1].

Images: This report makes use of images provided by DHS Office of Biometric Identity Management in May 2019. That collection is comprised of images and limited metadata indicating in which operation the data was collected e.g. airport-entry, pedestrian land entry, or exit. Some images were accompanied by metadata including date of capture, year of birth, sex, and country-of-birth[1]. From that database, this report uses 132 931 EXIT images of 128 384 individuals to search 567 air-ENTRY galleries each represented a departing flight[2]. Those galleries hold images drawn from the 825 976 airport ENTRY images of the 122 387 EXIT individuals who have a prior ENTRY image. The EXIT images were collected in 2018 and the first four months of 2019.

Other content: Section 1 discusses more general error sources and metrics relevant to EXIT and departure, putting matching results into the broader context of aircraft boarding. Section 2 guides readers toward different testing methodologies appropriate to answering a broader range of questions. Section 3 details our simulations and results. Section 4 considers use of one-to-many traveler verification systems (TVS) with a much larger population of N = 42 000 enrollees for use at other airport touchpoints. Importantly, section 5 discusses various reasons that would render the accuracy estimates in this report too high or too low.

Biometric results: We show that as many as 428 of 567 simulated flights each carrying 420 passengers can be boarded using one-to-many face recognition without any false negative errors - see Table 1 column 5. Stated in terms of error rates, this corresponds to at least 99.5% of travelers being able to board with a single presentation to a camera. This is attainable by enrolling a single prior ENTRY image in the galleries and using any of seven 2020-2021 face recognition algorithms - see Table 2 column 5.

For many travelers, multiple prior images can be enrolled in a gallery. Here, if we enroll an average of six prior air-ENTRY images, then the most accurate algorithm will now board 545 of 567 flights without any errors - see Table 1 column 4. Large gains are realized by all algorithms: Now at least 18 developers' algorithms are effective at boarding greater than 99.5% of travelers - see Table 2 column 3.

In 2007, U.S. legislation[3] specified that 97% of travelers' exits should be verified. That requirement can be met with almost all of the algorithms tested here[4]. Note that there are various systematic reasons why such accuracy may not be achieved in practice - see section 5.

In test of late 2018 algorithms [1], the most accurate algorithms on large population mugshot searches were NEC-2 and NEC-3. They remain in the top five on that benchmark today. However, when matching lower quality EXIT to ENTRY images the algorithms are less accurate than a 2018 Microsoft algorithm and many other more recent algorithms. By taking 100 minus the miss percentages in Table 2, NEC-3 correctly identifies 98.7% of individuals enrolled with a single image and 99.0% of those enrolled with images from multiple prior encounters. For the most accurate algorithm, Visionlabs-10, these values are 99.9% and 100% respectively, corresponding to about a factor of 10 fewer errors than NEC-3. Note that NEC-3 is now more than two years old and we may assume NEC has since improved its capability.

[1] This metadata was vital to our 2019 quantification of demographic effects in NIST Interagency Report 8280. [2]

[2] The terms ENTRY and EXIT refer respectively to inbound and outbound border crossings to, in this case, the United States.

[3] See 8 U.S.C. 1187(c).

[4] We chose to run only recent and high-performing algorithms and also some widely used prior-generation algorithms. Many more algorithms have been entered into the 1:N search track of FRVT.

			NUM ZERO FALSE NEGATIVE SIMULATIONS		
ALGORITHM			$N = 420$	$N = 420$	$N = 42000$
#	NAME	DATE	$k \geq 1$	$k = 1$	$k = 1$
1	VISIONLABS-010	2021-02-05	[1]545	[1]428	[3]177
2	IDEMIA-008	2021-03-15	[2]536	[2]422	[2]215
3	VISIONLABS-009	2020-08-04	[3]533	[3]406	[5]125
4	CLOUDWALK-HR-000	2021-02-10	[4]528	[4]393	[1]265
5	DEEPGLINT-001	2020-07-23	[5]519	[5]336	[4]153
6	CANON-CIB-000	2020-10-19	[6]518	[7]307	[13]19
7	XFORWARDAI-001	2021-01-21	[7]513	[6]309	[7]113
8	PARAVISION-007	2021-02-01	[8]490	[8]237	[6]124
9	TRUEFACE-000	2021-01-27	[9]476	[15]154	[15]4
10	NEUROTECHNOLOGY-008	2021-03-26	[10]470	[12]169	[17]2
11	COGENT-004	2021-02-10	[11]454	[11]182	[14]10
12	PARAVISION-005	2019-12-11	[12]453	[13]156	[10]72
13	NTECHLAB-008	2020-01-06	[13]451	[17]125	[20]1
14	PIXELALL-004	2020-07-02	[14]435	[16]146	[23]0
15	TECH5-002	2021-04-07	[15]416	[18]110	[18]2
16	DERMALOG-008	2021-01-25	[16]382	[20]71	[27]0
17	IDEMIA-007	2020-01-17	[17]374	[21]66	[21]0
18	MICROSOFT-006	2018-10-29	[18]361	[14]155	[16]3
19	SENSETIME-005	2020-12-17	[19]319	[9]233	[8]99
20	SENSETIME-004	2020-08-10	[20]316	[10]208	[9]96
21	RANKONE-010	2020-11-05	[21]300	[19]76	[22]0
22	RANKONE-009	2020-06-26	[22]203	[24]38	[19]1
23	COGNITEC-004	2021-03-08	[23]201	[26]11	[28]0
24	NEC-003	2018-10-30	[24]111	[22]66	[12]30
25	NEC-002	2018-10-30	[25]111	[23]65	[11]32
26	NEUROTECHNOLOGY-007	2019-10-03	[26]90	[25]21	[25]0
27	DERMALOG-007	2020-02-12	[27]30	[27]3	
28	IDEMIA-004	2018-06-30	[28]3	[29]0	[24]0
29	NEC-000	2018-06-21	[29]0	[28]0	[26]0

Table 1: **Number of simulations (out of 567) completed without errors.** *The second row N values give the number of individuals enrolled in each gallery. The 420 person galleries represent aircraft boarding; the 42000 case represents a airport security line where many more people are expected. The third row k values give the number of images of each enrollee in each gallery. The second and third columns identify the algorithm and the date it was submitted to NIST. The remaining columns give the number of simulations, out of 567, for which all 420 travelers boarded the flight (cols. 4, 5), or passed the checkpoint (column 6), without experiencing a false negative. Higher values are better, and the table is sorted on the first results column. The threshold is set so that only a fraction, 0.0003, of non-mated searches would return any match. The shaded cells indicate the three most accurate algorithms for that trial.*

Our demonstration of considerably higher accuracy from newer algorithms is an existence proof that EXIT accuracy on operational images can be improved. Given the pace of developments associated with the industrial migration to various convolutional neural networks, it is incumbent on end-users to establish contractual provisions for technology refreshment, factoring in such quantities as speed, scalability, stability, and cost.

The accuracy values noted above correspond to correct identification of an individual – here "correct" requires the algorithm to report the correct identity with a score above a set threshold. The threshold is set to limit false positives – this is necessary to prevent illicit boarding of an aircraft in an access-control context, and to limit visa-holder's status indicator mistakes in an EXIT facilitation context. The false positive identification rate (FPIR) in this report is usually set to 1 in 3 333, i.e. the proportion of searches of people not entitled to board an aircraft who succeed in doing so. A false positive occurs when a photo from such a traveler matches any (random) gallery photo. The consequences of such events, and a more detailed discussion of security, appears in section 1.5. We also include figures showing the tradeoff of false negative and positive identification rates, noting that some algorithms can afford lower FPIR without greatly degrading accuracy. An FPIR of 1 in 3 333 would imply that a mismatch would occur once during the boarding of about eight flights (3 333/420) – whether that is too frequent or to scarce is essentially policy issue informed by the error tradeoff characteristics of section 3.2.3 and the demographic dependencies given in sections 3.2.4 and 3.2.5.

Discussion: The report documents accuracy or small-gallery identification simulations showing a strong algorithm effect - accuracy is much improved with some algorithms versus others. This dominates two other main effects - first

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

ALGORITHM			PERCENT TRAVELERS NOT MATCHED		
			$N = 420$	$N = 420$	$N = 42000$
#	NAME	DATE	$k \geq 1$	$k = 1$	$k = 1$
1	VISIONLABS-010	2021-02-05	[1]0.02	[1]0.13	[3]0.61
2	IDEMIA-008	2021-03-15	[2]0.02	[2]0.15	[2]0.49
3	VISIONLABS-009	2020-08-04	[3]0.03	[3]0.16	[5]0.74
4	CLOUDWALK-HR-000	2021-02-10	[4]0.03	[4]0.18	[1]0.43
5	DEEPGLINT-001	2020-07-23	[5]0.04	[5]0.24	[6]0.80
6	CANON-CIB-000	2020-10-19	[6]0.04	[7]0.30	[13]1.79
7	XFORWARDAI-001	2021-01-21	[7]0.05	[6]0.28	[7]0.81
8	PARAVISION-007	2021-02-01	[8]0.07	[8]0.41	[4]0.72
9	TRUEFACE-000	2021-01-27	[9]0.08	[14]0.66	[16]3.72
10	NEUROTECHNOLOGY-008	2021-03-26	[10]0.08	[11]0.59	[18]4.12
11	PARAVISION-005	2019-12-11	[11]0.10	[13]0.62	[10]1.04
12	NTECHLAB-008	2020-01-06	[12]0.11	[17]0.81	[19]4.52
13	COGENT-004	2021-02-10	[13]0.11	[12]0.59	[14]2.34
14	PIXELALL-004	2020-07-02	[14]0.12	[15]0.69	[17]3.88
15	TECH5-002	2021-04-07	[15]0.14	[18]0.86	[21]5.21
16	DERMALOG-008	2021-01-25	[16]0.19	[19]1.04	[23]6.39
17	IDEMIA-007	2020-01-17	[17]0.19	[21]1.12	[20]5.19
18	MICROSOFT-006	2018-10-29	[18]0.23	[16]0.71	[15]3.21
19	SENSETIME-005	2020-12-17	[19]0.28	[9]0.45	[8]0.85
20	SENSETIME-004	2020-08-10	[20]0.29	[10]0.50	[9]0.89
21	RANKONE-010	2020-11-05	[21]0.31	[20]1.06	[22]5.71
22	COGNITEC-004	2021-03-08	[22]0.49	[26]2.18	[25]9.20
23	RANKONE-009	2020-06-26	[23]0.52	[24]1.52	[24]7.85
24	NEC-002	2018-10-30	[24]0.99	[22]1.29	[11]1.61
25	NEC-003	2018-10-30	[25]0.99	[23]1.29	[12]1.78
26	NEUROTECHNOLOGY-007	2019-10-03	[26]1.02	[25]2.02	[27]31.93
27	DERMALOG-007	2020-02-12	[27]1.97	[27]3.66	
28	IDEMIA-004	2018-06-30	[28]4.96	[28]8.13	[26]17.81
29	NEC-000	2018-06-21	[29]15.41	[29]18.85	[28]91.97

Table 2: **False negative rates by gallery size and number of enrolled images per person.** *The second row N values give the number of individuals enrolled in each gallery. The 420 person galleries represent aircraft boarding; the 42000 case represents a airport security line where many more people are expected. The third row k values give the number of images of each enrollee in each gallery. The second and third columns identify the algorithm and the date it was submitted to NIST. The remaining columns give false negative identification "miss" rates i.e. the proportion of travelers not matched to their gallery photo(s), expressed as a percentage. Lower values are better, and the table is sorted on the first results column. The superscripts give the rank of the algorithm for that column. The threshold is set so that only a fraction, 0.0003, of non-mated searches would return any match. The shaded cells indicate the three most accurate algorithms for that trial.*

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

that more prior enrollment images for each enrollee improves accuracy and, second, that even a 100-fold population size increase degrades accuracy only modestly.

The report gives some information on demographic dependencies. Many algorithms give somewhat higher false negative rates on women compared to men. This is not true for, or has reduced magnitude, for the more accurate algorithms. With high accuracy, and with opportunities in real operations to make second identification attempts, these differentials are either small or can be remediated. The report also notes demographic dependence on false positive rates, particularly that women and people of certain nationalities, often East Asia, tend to give higher false positive identification rates. Again some algorithms are considerably superior to others in this respect. Note that security context matters: In particular that passive non-mate, and active attack, presentations will be very small percentages of all attempts.

The accuracy estimates in this report are just that, estimates. Section 5 notes several factors that would drive accuracy higher or lower. Primary among those is that we can't be sure how well the images we possess represent the actual paired ENTRY galleries and their EXIT photos. A passport control authority has two complementary options for improving on our estimates: First is to run exhaustive clipboard style operational tests; second is to provide NIST or some other laboratory with a) actual images, and b) the operational algorithm. This latter option had been planned in 2019 but was derailed for several reasons, including the COVID pandemic.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

1 Errors and Their Consequences in Biometric Exit

The following subsection describe mechanisms by which an EXIT system, as comprised, makes errors. We distinguish biometric errors (from cameras and algorithms) from operational issues deriving from business processes.

1.1 Failure to Enroll

Nature: In the context of TVS' manifest-driven gallery construction, some individuals who are legitimately booked on an aircraft will not be enrolled in the face recognition gallery. This number will usually be zero but could be non-zero for several reasons, among them:

1. Absence of historical photo. For various policy-related issues a PCA may not have a prior photo - these could include first-time visitors, foreign passport holders born in the country, and bilateral trade-related visa exemptions. In such cases a PCA might legitimately have no ENTRY record. This circumstance might be termed an *operational failure to enroll*.

   Measurement: A PCA can estimate the prevalence of missing enrollments by cross-referencing airline manifests and the lack of prior reference photos. This estimate will include instances of 2 below.

   Consequences: Failures to enroll will manifest as false negatives (see section 1.4 below). Airline staff can resolve by biographic and human visual biometric inspection.

2. Biographic errors. It is possible that the manifest provided to the PCA by the air carriers includes biographic errors from well understood sources such as recent marriage and change of name, and typographical errors.

   Measurement: A PCA can estimate the prevalence of missing enrollments by cross-referencing airline manifests and the lack of prior reference photos. This estimate will include instances of 1 above.

   Consequences: Failures to enroll will manifest as false negatives (see section 1.4 below). Airline staff can resolve by biographic and human visual biometric inspection.

3. Poor image quality. It is possible the photographs that a PCA has on an individual are of poor enough quality that the TVS feature extraction software fails to produce a template from the photograph. This could occur because the face detector fails to find the face, or because the software deems the photo to be of low utility to their downstream recognition engine so, electively, does not produce a template. Such outcomes would constitute *biometric failures to enroll*.

   Measurement: A PCA can estimate algorithm enrollment failures by direct analysis of TVS logs.

   Consequences: Failures to enroll will manifest as false negatives (see section 1.4 below). Airline staff can resolve by biographic and human visual biometric inspection.

1.2 Failure to Capture

Nature: During aircraft boarding TVS never receives photos of some travelers for at least two reasons:

1. Camera failure: Some cameras might fail to trigger and take a photograph. This can occur due to failed face detection (e.g. due to sunglasses, or subject not being in the field-of-view), or because an on-board quality algorithm deemed the captured photograph of insufficient utility, or due to some system fault of the kind remedied

by rebooting the system. During observations at various airports in June 2019, some cameras would not trigger; others would trigger only after the subject disengaged by moving away, and then re-engaged.

Measurement: We can put an upper bound on the frequency of such events by subtracting the number of people verified from the number of people on the manifest. This quantity will include outright recognition failures too. This estimate will include people who never appeared before the camera (e.g. because the airline allowed traditional paper-based boarding).

2. Airline operations: An operational source of "failure to capture" can be that airline staff might redirect the traveler to some human-adjudicated boarding process such as the traditional passport or boarding-pass based biographic confirmation. This could occur a) because the staff perceive the traveler has had difficulty, or b) that they will have difficulty (e.g. because they're too tall or short), or c) simply because the airline staff are trying to expedite boarding by using the biometric process and the biographic process.

   Measurement: Such events can only be documented by observation, most readily human observation, but also via some automated supervisor or logging system.

Consequences: For an in-scope traveler the consequence will be that EXIT will only be recorded biographically according to the information used in forming the passenger manifest – this is essentially the legacy biographic process. An immediate operational consequence is that the passenger will have to be processed manually (by airline) staff. Downstream, this may cause the PCA to perform overstay inquiries.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

1.3 Failure to Extract Features

Nature: It is possible the photographs that the PCA has on an individual are of poor enough quality that the TVS feature extraction software fails to produce a template from the photograph. This can occur during gallery construction or during EXIT operations.

Measurement: Such events can be measured from algorithm logs such as those produced in FRVT, and likely by operational systems.

Consequences: If TVS fails to extract features during EXIT, the traveler's boarding attempt will be rejected, possibly silently. He or she may make a second attempt, perhaps after being prompted. In June 2019 observation of boarding, the author noticed airline staff directing passengers to the gate-agent biographic process. This would likely lead to the PCA having to revert to its reliance on biographic recording of EXIT.

1.4 False Negative During Identification

Nature: In a positive identification application like EXIT, the one-to-many search algorithm generally grants access if the rank-one (i.e. highest-scoring) candidate has a score above threshold. The identity of the person in the live photo is taken to be that returned by the system even if it is incorrect. From a testing perspective, an error occurs if the rank-1 candidate is of the wrong identity or has score below threshold. This gives us the following performance metric, the false negative identification rate (FNIR):

$$\text{FNIR}(N,T) = \frac{\text{Num. searches where top-scoring candidate has wrong ID or score below threshold}}{\text{Number of searches conducted}} \tag{1}$$

This definition automatically incorporates "failure to extract" feature events as they won't return high-scoring candidates. The dependence on gallery size, N, and threshold, T, are present as they are design choices affecting FNIR. For an audience who likes to think in terms of accuracy or "hit rates", we can convert the "miss rate" or Eq. 1 to True Positive Identification Rate using TPIR = 1 – FNIR, so a 3% FNIR becomes 97% TPIR. However, that definition is naïve in that it assumes every traveler was photographed. It ignores instances of failure-to-capture, and also cases where travelers are photographed, not matched, and then make further attempts. Then an operational definition of false negative identification rate is

$$\text{FNIR}(N, T) = \frac{\text{Num. travelers who are not matched to the correct ID in one or more presentations to the camera}}{\text{Number of travelers}} \quad (2)$$

The two measures would be equivalent if each traveler executes just once search. To the extent that is true, our Equation 1 estimates in this report will approximate Equation 2. We use the 1 thoughout this report. We discuss in section 5, factors that can make our estimates too high or too low.

Measurement: In this report, we don't have insight into the transactional nature of aircraft boarding, with failed captures or failed searches. Instead all we see are images that can be used in simulations of boarding. For measuring duration of boarding, and quantities such as the number of travelers who need to make further presentations, an operational observational test is most appropriate. Many aspects may be measurable in scenario tests in which passengers and airline staff model the actual target boarding process.

Consequences: False negatives will usually be resolved by biographic and human visual biometric inspection by airline staff. For an in-scope traveler the consequence will be that EXIT will only be recorded biographically according to the information used in forming the passenger manifest – this is essentially the legacy biographic process. Downstream, this may cause the PCA to perform overstay inquiries. The PCA would possess an aircraft boarding photo, but one that is not bound to an identity – such images are provided to PCA staff monitoring a flight departure.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

1.5 False Positive During Identification

Nature: False positives occur when images of two people are erroneously associated. In biometric EXIT there are three kinds of false positive:

- First is the **in-gallery false positive** in which a legitimately enrolled traveler matches the wrong identity. Such a possibility necessarily implies that the correct identity would be displaced from the rank-1 position on the candidate list, usually to rank 2. That list is a data structure internal to the particular TVS and is not typically presented to airline staff or anyone else. Depending on how the system is built, an in-gallery false positive may result in a false negative for the correct passenger if he or she boards later in the process. Such errors were observed by the author in June 2019 during visits to observe the boarding process in five different airports.

  Measurement: The in-gallery false positive rate is not currently defined in performance testing standards as it is approximately the proportion of mated searches yielding the mate at rank 2 or higher. Such an outcome would most often occur because the search imagery is of poor quality, but could occur if the enrolled imagery was poor. Formal measurement can be achieved by careful online observation of the boarding process. Error rates can be estimated approximately from recognition logs by counting instances of a passenger apparently boarding the plane twice – once legitimately as themselves and secondly when another traveler incorrectly matched their identity.

  Consequence: Such errors will likely be resolved by airline staff, who may become familiar with such an event.

- Second is the **incorrect acceptance** of people who are not in the gallery and not expected on the departing flight.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

This population includes travelers who mistakenly arrive at the wrong gate[5] without subversive intent. The frequency of occurrence is usually stated by the False Positive Identification Rate (FPIR). FPIR is the primary security-related parameter in a one-to-many access control system. Its value is chosen by a system owner to target security objectives and is implemented by setting the system threshold according to some calibration[6].

Measurement: Such errors were observed by the author in June 2019 when airline staff in the gate area were accidentally captured by the camera and incorrectly matched to an actual passenger. While this kind of error could be measured by making in-person attempts, this approach does not scale. An offline approach in which images are matched after-the-fact affords more precise FPIR estimates – this report takes just this approach.

Consequence: The consequence for the airline is potentially a stowaway. However, airlines usually count passenger totals and may thereby be able to detect such events. While there is little consequence for the PCA's EXIT processing, these events, if undetected, could cause erroneous updates to the PCA's systems, undermining integrity.

- Third is a false positive from someone who is illicitly trying to gain access. This category would include stowaways and potentially visa overstayers.

Passive vs. active attack

False match rates usually express the likelihood that a face recognition algorithm will compare two photographs and return a high score from two individuals who are selected entirely randomly, or perhaps with the restriction that they have the same demographics such as age, sex, and race.

However, if someone makes more deliberate efforts to impersonate an identity e.g. via cosmetics or wearing a face mask, then additional algorithms must be employed to detect the presentation attack (PA). To succeed an attacker must defeat the PAD subsystem, if installed and enabled, AND match the intended identity – see section 1.6

- **Casual attack:** If someone is making a low-effort attack – for example as a stowaway – they might rely on matching any identity essentially fortuitously, and then hoping the airline staff does not notice nor take steps to resolve the match. A second intent here would be to fake someone's departure from a country. This possibility - to overstay a visa by sending a confederate to verify a particular identity - is notable in that it would be difficult for an overstayer to select a confederate who would match the *particular* identity in a biometric search – In this respect a one-to-many system where there is no claim to an identity is more secure to passive attack. However, the security context is that such a system is prone to circumvention attack: a confederate failing to match an enrolled identity might appeal to airline staff who would make biographic or visual biometric efforts to verify the person, with the likely outcome that passenger would be allowed to board.
- **Active attack:** An overstay attempt would be much more successful if the confederate actively impersonates the visa-holder. This could be achieved using a presentation attack instrument such as a face mask.

  Measurement: Vulnerability to active attack could be demonstrated via "red-team" presentations to the operational system. More formal quantification of the vulnerabilities is best conducted in laboratory trials using identical equipment to that used in the operation. Each approach will require controlled, defined and

[5] This can occur because of a gate change, or because someone goes to the wrong gate. The author, for example, has accidentally tried paper-based boarding at the adjacent gate on several occasions.

[6] Threshold calibration is an imprecise process because FPIR often depends on demographics and image quality related properties. A threshold is set starting with vendor recommendation and refined using offline tests (such as FRVT) or empirical instrumentation and tests or logging of the operational system.

repeatable production of presentation artefacts (masks, cosmetics etc.). The metrics relevant to this kind of attack are standardized – see section 2.2.

- **Comparison with existing paper-based boarding:** Attacks on non-biometric paper-based departure systems are possible also: A stowaway could find, or steal, a boarding pass. A confederate seeking to depart for a visa-overstayer would only have to present a boarding pass and possibly a cursory inspection by the airline staff of the passport. In these cases, a biometric system, if used and not circumvented, will improve security compared over the legacy process.
- **Consequences:** For an IA, a successful impersonation attack would likely produce an undetected overstay. The attack assumes the confederate either does not need or want to return to the United States or could do so using other documents. There are no consequences for the airline.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

1.6 Presentation Attack Detection Metrics

The ISO/IEC 30107-3 standard establishes the metric Impostor Attack Presentation Match rate (IAPMR) which expresses the proportion of attackers who both defeat the PA detection software AND match the correct identity. That metric is appropriate to access, say, to a mobile phone. In one-to-many processing such as paper-less EXIT, a traveler would have to defeat the PAD and match the specific intended enrollment.

1.7 Demographic Differentials

Biometrics generally give different error rates for different populations. For example, fingerprints are known to give higher false negative rates in the very young and the elderly[7]. NIST Interagency Report 8280 [2] documented error rate differentials for face recognition examining the effect of sex, age and race on accuracy of many commercial algorithms. That report made an important distinction between differentials in false negative and false positive error rates, the former affecting how well a single individual is not matched as him or herself, the latter affecting how often two individuals are erroneously associated. The consequences of such errors, and differentials in their rate of occurrence, are very different. We include visualizations of false negative differentials in section 3.2.1 and false positives demographic differentials in sections 3.2.4 and 3.2.5.

[7] In the young, typical contact sensors have inadequate resolution to resolve the fine friction ridge structure. In the elderly the factors include inelasticity of the skin and inability to present flat impressions e.g. due to arthritis.

2 Operational Questions

2.1 Context

This report gives extensive documentation of biometric identification performance. However larger questions exist, and core biometric performance statements only inform answers to those questions. For example,

- An airline might ask "which camera and boarding solution should we procure?" – this report is silent on that because we would at least need to know what cameras were used for collecting the data, and this is not information we have. Dedicated laboratory tests of camera equipment[8] are appropriate to such tests.
- An airline might ask "what is the proportion of passengers being referred to gate agents" – such a quantity could be approximately estimated from TVS logs, but is more precisely answered only by observation of the operational system.
- A security analyst might ask "what is the chance on an active impersonation attack succeeding" – this question can be addressed potentially by laboratory trials if the fielded system can be copied and if access access to the TVS recognition engine is granted. It may be easier to conduct operational "red team" trials with an appropriately motivated staff. Active attacks (e.g. using face masks) are not the fault of the recognition algorithm per-se, but are enabled by lack of (or use of poor) presentation attack detection algorithms[9].
- A policy maker might ask "is biometrics better than biographic matching for overstay detection?" – we can't address that without biographic data and extant biographic matching algorithms.

2.2 Standardized Tests

Since 2003, there have been significant worldwide investments in supporting development of biometrics performance testing and reporting standards in the ISO/IEC JTC 1 Subcommittee 37. That body develops very well vetted consensus standards in working groups (WGs) dedicated to vocabulary (WG1), interfaces (WG2), data interchange and image quality (WG3), application aspects including face-aware capture devices (WG4), performance testing and reporting (WG5) and societal issues (WG6). Table 3 lists standards that may be valuable in the measurement of performance in a PCA's ENTRY-EXIT processes.

There are a number of other testing standards supporting other domains of use.

This publication is available free of charge from: https://doi.org/10.6028/NIST IR 8381

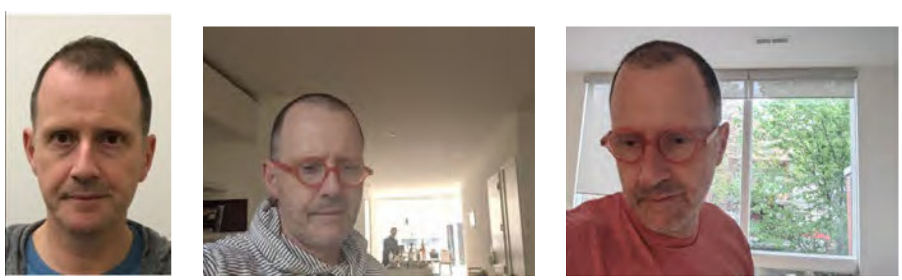
[8] See the scenario tests conducted at the Maryland Test Facility, for example.

[9] PAD approaches have advanced in recent years, both in software and hardware. However, their use will often increases false negatives because they sometimes erroneously flag a bona-fide presentation. Their use may be more appropriate on inbound arrival processing (ENTRY).

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

Table 3: Testing standards supporting performance measurement in ENTRY-EXIT

Number	Title	Relevance
ISO/IEC 19795-1	Principles and Framework	This foundational document establishes requirements on all biometric tests regarding design of tests of enrollment, verification and identification, and how to put uncertainty estimates on measured error rates.
ISO/IEC 19795-2	Technology and Scenario Testing	Regulates two kinds of "in-vitro" test: "Technology" tests which are most often offline sample comparison and search tests such as those documented herein, and "scenario tests" that are usually human-in-the-loop laboratory tests intended to mimic operational systems.
ISO/IEC 19795-3	Environmental Aspects	A technical report guiding testing and reporting in the presence of environmental variations such as humidity and illumination
ISO/IEC 19794-4	Interoperability Testing	Relevant to tests where components of a system, possibly from different manufacturers must produce and consume standardized data, for example cameras must produce images that will be consumed by remote recognition algorithms.
ISO/IEC 19795-6	Operational Testing	Establishes requirements on "in-situ" tests, where identity ground truth is not necessarily known, and where the act of measuring accuracy or duration can potentially disturb the estimates. This kind of test is advantaged by considering the actual system on its native population in its native environment. These aspects are often material and difficult to approximate in lab tests.
ISO/IEC 30107-3	Presentation Attack Detection	This standard regulates tests of PAD components and PAD-enabled systems and gives detailed guidance on measuring and naming of error rates that are available for various levels of logging and instrumentation.
ISO/IEC 19795-10	Demographic dependence	This standard (2020-11) is in the early stages of development. It will establish requirements on various kinds of tests intended to measure demographic differentials in biometric devices, algorithm and systems.



a) Visa or Passport | b) Immigration Lane ENTRY | Departure Point EXIT

Figure 1: Image (a) is representative of passport-like data that would ordinarily be available to a PCA's TVS from all in-scope travelers and citizens. However, such images were not available for the trials conducted here. The remaining images have size 240x240 pixels and are representative of some poorer quality ENTRY images: Image (b) is typical of ENTRY photos in that it has has non-frontal pose, and strong background illumination reducing contrast on the subject's face. Image (c) is typical of EXIT photos in that it exhibits some close-range distortion, mild non-frontal pose arising from "don't wait for frontal presentation" fast-capture and adverse background lighting. contrast.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

3 Simulation and Results

3.1 Air-Exit Simulations

We simulate biometric EXIT by running simulations using archived images as follows.

1. We form a departing flight by placing ENTRY images from N = 420 individuals into a gallery. We use 420 because that number is reasonable for a large commercial twin-aisle jet such as the Boeing 777 or the Airbus A380[10]. The exact gallery size is not that important because accuracy is an insensitive function of N. We later increase the population size to 42 000 to simulate an airport security checkpoint, for example.

2. We populate an EXIT gallery in two ways.

   (a) First, with one ENTRY image per person.

   (b) Second with a multiple such images, the average is about 6, with some variance per individual.

   While it is common practice to populate the gallery with images from all prior encounters of a person[11], we include the one-image case to show "worst-case" accuracy i.e. that expected when only one prior encounter is available. We include results for single- and multiple-image enrollment in sections 3.2.1 and 3.2.2 respectively[12].

3. We populate a gallery with individuals from the same region of the world. We do this for two reasons: As discussed in section 5, we list various factors that will push our error rate estimates up, and down. that flights departing the U.S. tend to have some racial homogeneity – flights departing for Japan have more individuals from East Asian countries than do flights departing to Nigeria, and more than would be expected by random selection. Another reason is that face recognition accuracy will be worse for homogenous galleries because false positives will be more common. Our practice of building homogenous galleries biases the test toward higher error.

4. The 12 regions are: Europe, W. Africa, E. Africa, N. Africa, Middle East, S. Asia, E. Asia, Oceania, N. America, C. America, Caribbean and S. America. We assign individuals to a region based on the issuers of their travel document. Occasionally some travelers will travel on a different country's travel document; in such cases we assume their region to be that of the gallery ENTRY image.

5. We form 567 galleries, with one image per person. We form another set of 567 galleries with variable numbers of images per person. The number of galleries we can form per region varies because we have more images from some regions than others.

6. We search each gallery using a single probe-set containing 127 258 EXIT images of 123 075 people. By visual inspection it is evident that the images are collected using different cameras in different locations. For a given gallery only small proportion of the searches will have an enrolled mate in the departure gallery, at most 420 of 123 075 people. These mated pairs afford estimates of false negative identification rate. The remaining images, from persons of all regions of the world, form a non-mated search set used for estimating false positive identification rates.

[10] Aircraft configuration makes a difference, so that while the A380 is capable of carrying 560 economy class passengers it is atypical for that to occur for aircraft departing the United States.

[11] Not all, as the U.S. PCA stated, their TVS "does not enroll recent crossing images of U.S. travelers into the gallery, but does enroll recent crossing images of foreign nationals into the gallery."

[12] The single-image enrollment will be more pertinent to processing of citizens of a country for whom, often, only one photo exists in the gallery. The multiple-image enrollments yield better accuracy, and are pertinent to foreign travelers.

7. We run multiple algorithms, in some cases more than one from each developer. These were submitted to the one-to-many identification track of the FRVT between May 2018 and the present. The list of algorithms includes the NEC-3 algorithm that was broadly the most accurate through November 2018 as reported in NIST Interagency Report 8271 [1], but which has been eclipsed in accuracy by newer algorithms submitted since.

8. We compute 10 thresholds for each algorithm corresponding respectively to the 10 false positive identification error rates: 0.00003, 0.0001, 0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1. We get the threshold value by looking at the highest non-mate score produced when running all non-mate searches against all galleries. Given, say, 126838 non-mate searches into each of 567 galleries, the threshold for FPIR = 0.0003 is taken to be the 126838 x 567 x 0.0003 = 21575-th highest observed rank-one comparison score.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

3.2 Results

3.2.1 Attainable accuracy with single entry image

Figure 2 shows accuracy for two algorithms submitted to NIST 28 months apart. These are the NEC-3 algorithm submitted to NIST in November 2018, and the Visionlabs-10 algorithm from February 2021. The gallery size is N = 420 subjects, each person enrolled with exactly one ENTRY image. The vertical axis is a count of the individuals who are not biometrically authenticated during boarding. The horizontal axis shows the region of the enrolled population. The dots correspond of one departing flight. The dots are jittered horizontally around the region label, and vertically around the integer value, to avoid over-plotting and show the distribution.

The notable observations from the graphs are:

1. The number of false negative recognition errors is spread between zero and 16, with the most common value being 6. These errors would need to be resolved via a second attempt at biometrics, or via an airline-defined biographic process.

2. The distributions across regions are similar. The Central American flights give modestly higher FNIR, but this may simply be the result of chance. To the extent that some of the regions here are proxies for race, the results comport with those published in NIST Interagency Report 8280 [2] showing little dependence of false negative rates on race. Any false negative demographic differentials should be corrected for:

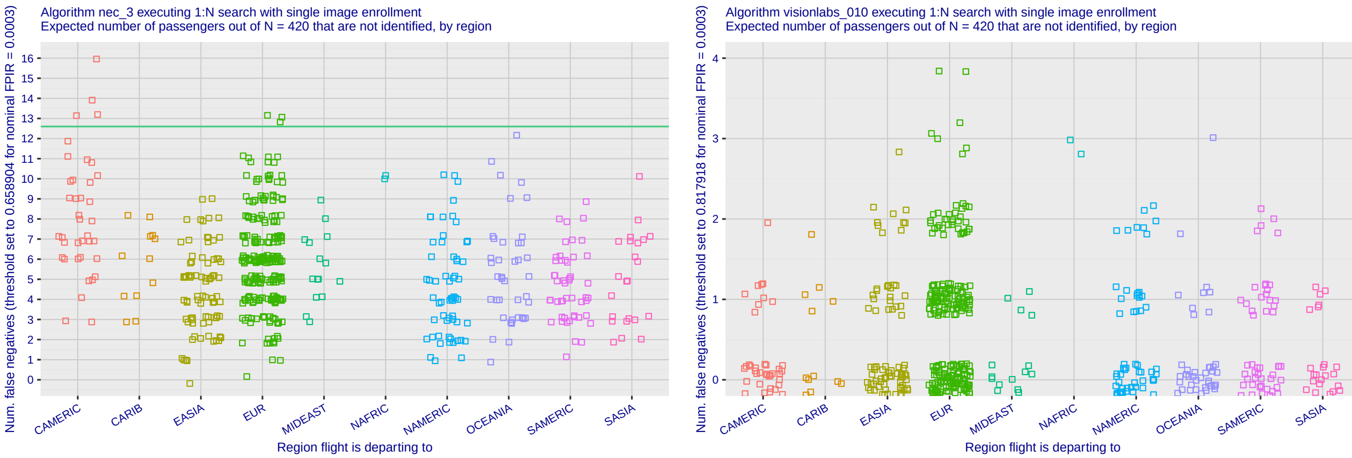
   (a) Ageing: It is possible that different travelers from certain regions travel less frequently such that the gallery photos are older – time lapse affects appearance and accuracy.

   (b) Age: It is possible that absolute age affects accuracy. For example, although not the subject of the simulation here, flights into Orlando are disproportionately populated with children[13] whose lower height can affect head pitch angle and accuracy.

3. We report the rate of false negatives (FNIR) in a subsequent figure – but note here that a count of 13 (i.e. 0.03 * 420) corresponds to a 3% failure rate. On that basis, the overall error rate is below 3% corresponding to better than the 97% verification rate required in 2007 legislation.

[13] In visits to observe EXIT boarding processes June 2019, the author observed children, without instruction, standing on tip-toes in order to present their face to the camera mounted above five feet. This was sometimes effective.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381



(a) NEC-3 (b) VISIONLABS-10

Figure 2: Count of false negatives on simulated flights by region using the NEC-3 algorithm from Nov. 2018 and the Visionlabs-10 version from February 2021. The gallery is populated with one ENTRY image from each of N = 420 individuals. The threshold is set to target a false positive identification rate of 0.0003 corresponding to 1 false positive in 3 333 impostor search attempts. The false negative identification rate for a flight can be stated by dividing the number of false negatives by the number of passengers, 420.

Figure 2(b) shows accuracy for a recent algorithm that is among the most accurate submissions to the one-to-many track of FRVT. The number of errors now is much lower, ranging from 0 to 4, with most common value being 0. A false negative count of zero corresponds to correct recognition of all passengers.

Note that the most accurate algorithms have been submitted to NIST recently, in early 2021, showing accuracy gains are still being realized by developer innovation. Several algorithms, including the VisionLabs-10 algorithm used in Figure 2(b), are more accurate than leading algorithms submitted to NIST in 2018 - see the ongoing FRVT webpage for names, dates, and more general accuracy results. The implication is that a PCA will realize accuracy gains if its technology refresh process is active and frequent.

Figure 3 shows the same figure for the most accurate algorithms tabulated appearing in Table 2. We note the following:

1. Figure 3 includes, in blue text, values for FNIR, the estimated proportion of passengers who will not be able to board with a single probe capture. The values are well near 0.1% for the most accurate algorithms, and often above 1% for the less accurate ones.

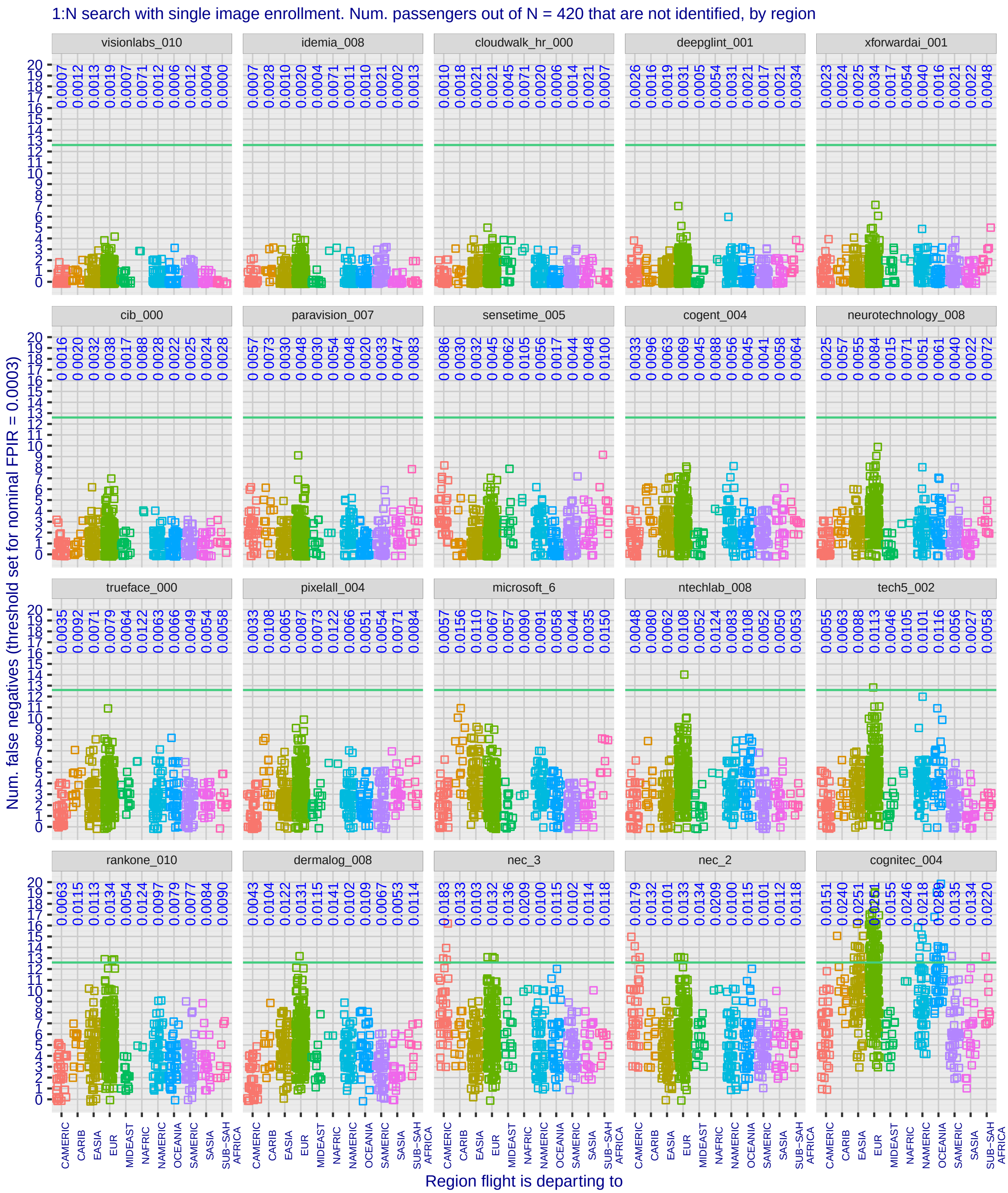


Figure 3: Count of false negatives on simulated flights by region. Each point corresponds to one flight with a gallery populated with one ENTRY image from each of N = 420 individuals. The threshold is set to target a false positive identification rate of 0.0003 corresponding to 1 false positive in 3 333 impostor search attempts. The blue text gives FNIR. The panels are arranged left-to-right, top-to-bottom in order of mean false negative count. The horizontal green line corresponds to the 3% false negative goal implied by legislation in the U.S.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

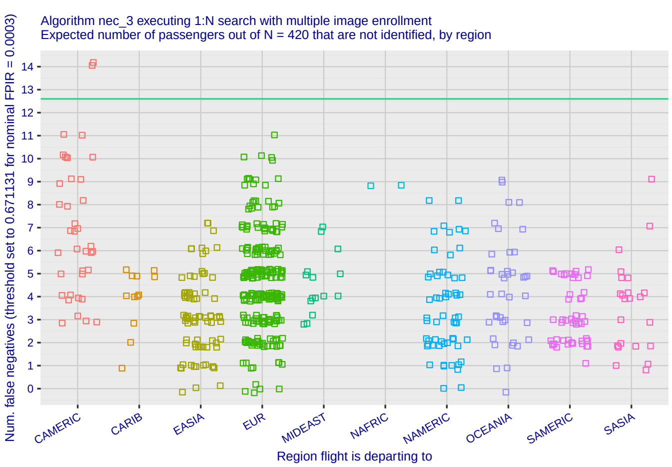
This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

3.2.2 Attainable accuracy with multiple entry images

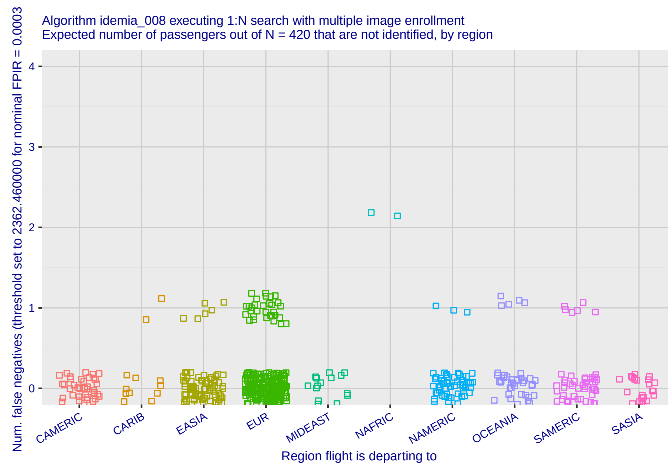
Figure 4 shows the accuracy results for two algorithms for a gallery of size N = 420 subjects each now enrolled with *multiple* ENTRY images. The algorithms were submitted 29 months apart, in November 2018 (NEC-3) and March 2021 (Idemia-8). From the two figures we note the following:

1. The use of multiple enrollment images reduces the number of false negative recognition errors modestly for NEC-3 (2018). It produces around 4 errors on average instead of 6 with a single image. The worst case count is reduced from 16 to 14.

2. With Idemia-8 (2021) the effect of enrolling more images is a more substantial reduction in false negative outcomes such that a large majority of flights will see all passengers board without any errors. The worst case count of error is reduced from 4 to 2.



(a) NEC-3



(b) IDEMIA-8

Figure 4: Count of false negatives on simulated flights by region using the November 2018 NEC-3 and March 2019 Idemia-8 algorithms. Each point corresponds to one flight the gallery for which is populated with multiple ENTRY image from each of N = 420 individuals. The threshold is set to target a false positive identification rate of 0.0003 corresponding to 1 false positive in 3 333 impostor search attempts. The horizontal green line corresponds to the 3% false negative goal implied by legislation in the U.S.

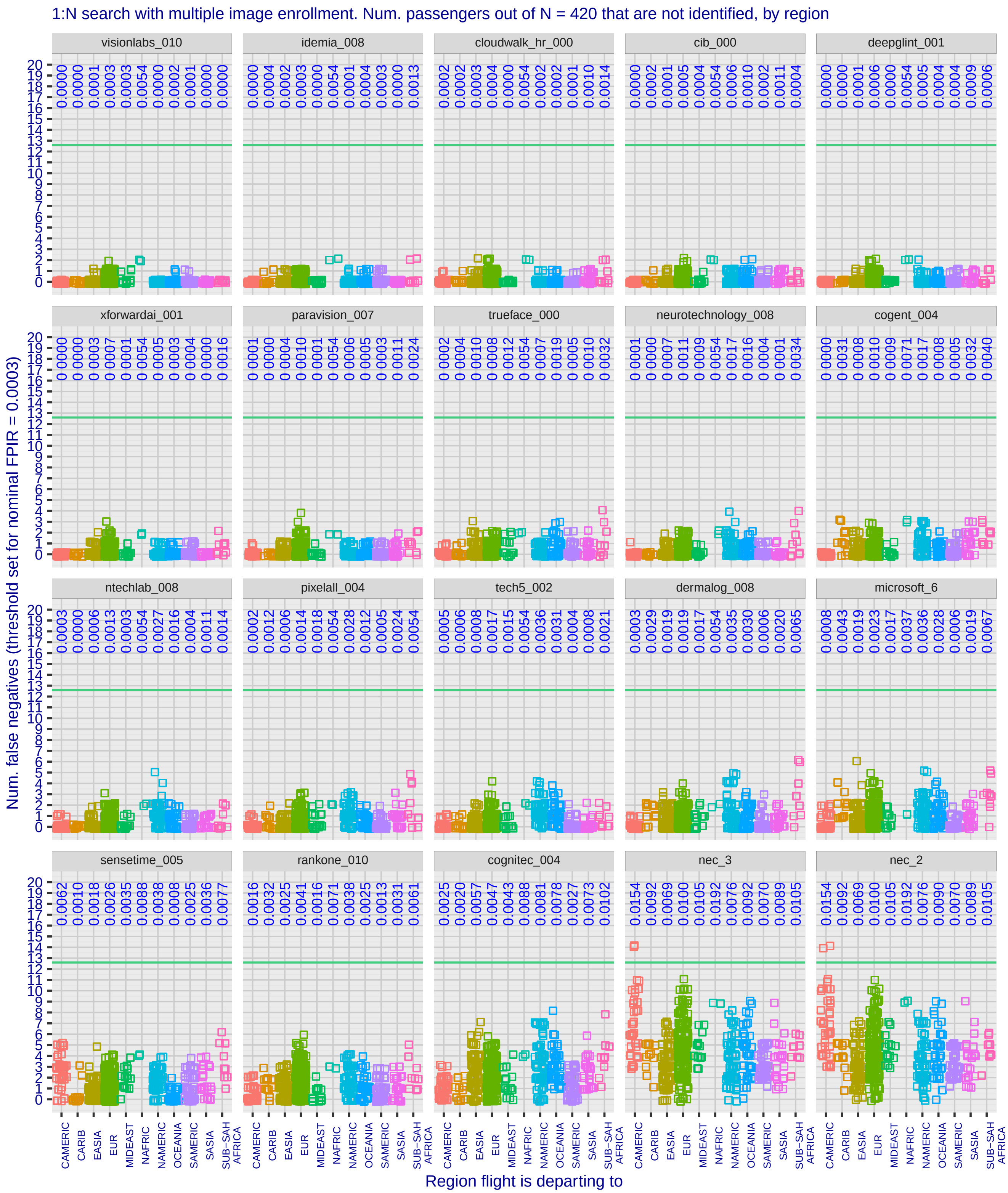


Figure 5: Comparison of false negative identification rates between number of images enrolled (one per person, vs. several) and between algorithms. The algorithms were submitted to between June 2018 and April 2021. Each point corresponds to one flight to the identified region the gallery for which is populated with ENTRY images from each of N = 420 individuals. The blue text is a false negative identification rate (FNIR), often below 1%. The orange text is the number of simulated flights, out of 567, for which the number of false negative errors is zero. The threshold is set to target a false positive identification rate of 0.0003 corresponding to 1 false positive in 3 333 impostor search attempts.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

The failure of NEC-3 to exploit multiple images may stem from how we provided images to the gallery. In FRVT we typically provide all images of an individual to the algorithm in one call to the template generation function – the algorithm consumes multiple images and has the opportunity to select or fuse images as it sees fit. However, that is atypical operationally: images are provided to the algorithm serially such that multiple images of the same person result in separate enrolled templates - that is the model followed here[14] even though it denies the algorithms an explicit fusion opportunity. Figure 5 shows analogous results for nine of the more accurate algorithms evaluated in FRVT through November 2011. The panels are included in order of mean overall number of false negatives. Notably:

1. For the majority of flights, the most accurate algorithms correctly identify every passenger, and only ever fail to on up to 2 out 420 people.

2. On this metric there are multiple algorithms affording lower false negative identification error rates than does NEC-3. This is an existence proof of better accuracy that suggests an PCA will benefit from monitoring of test results and regular technology refresh. A PCA would need to factor other variables into procurement from a new developer including performance aspects (speed, scalability to large galleries, and demographic equitability) and contractual factors like capital and transaction costs, including those of integration.

3.2.3 False negative vs. false positive tradeoff

The results for each algorithm thus far have been stated at a single threshold. If we had set this threshold to a higher value the false negative rates would also have been higher, but with the advantage of lower false positive rates. Conversely, if the threshold had been low, false negatives would be better and false positives could occur more easily. The threshold is conventionally set to achieve a low enough probability that an impostor could match an enrolled identity thereby meeting some planned security objective. In one-to-many applications, an impostor only needs to match any enrolled identity to gain access – he has N opportunities. This generally necessitates higher thresholds than one-to-one verification where the impostor claims one particular identity.

[14] See Figure 8 and section 3.2 in the FRVT 1:N report, NIST Interagency Report 8271 [1] for details on multi-image enrollment and metrics. See the FRVT page for newer algorithm results.

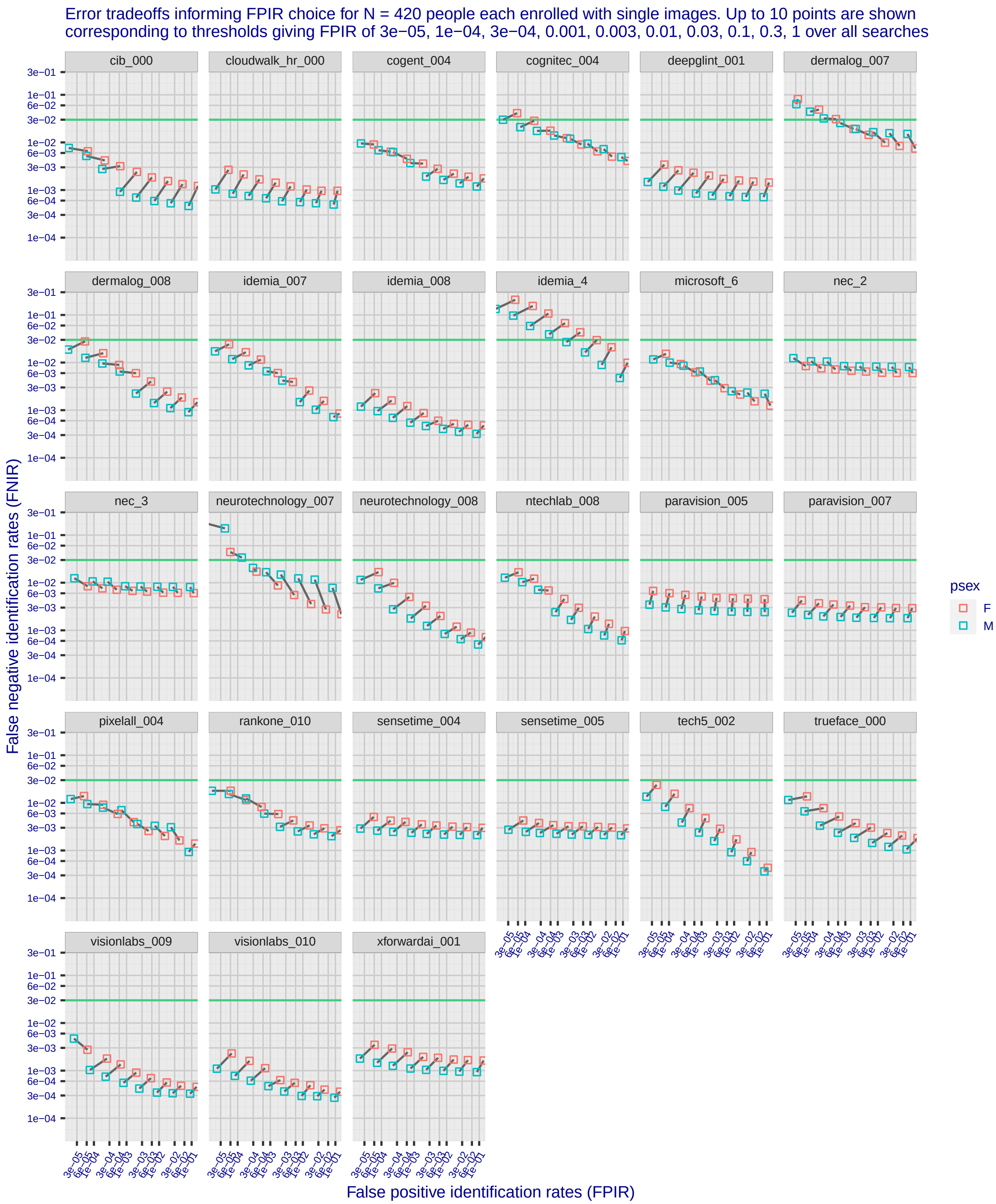


Figure 6: Error tradeoff for 26 algorithms executing 1:N searches with N = 420 people enrolled with a single image. The 10 point-pairs correspond to 10 possible thresholds for each algorithm. The red and blue boxes correspond to female and male travelers; their relative displacement indicates generally higher false positives and false negative rates in women. Smaller displacement indicates smaller (better) demographic differential (see NEC and Paravision-5, for example). The horizontal green line corresponds to 3% false negative goal implied by legislation in the U.S.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

What should the false positive identification rate be?

This question is about policy. As discussed in in the introduction a TVS can serve double duty as an aircraft access control system and as a visa-holder EXIT status facilitation system. The discussion centers on what the system is trying to prevent – stowaways, evasion of traveler-to-bag matching, or faking someone's departure.

One factor is the prior probability that someone would try to board a flight at all. It's likely quite common and not necessarily nefarious - the author has accidentally gone to the wrong gate on more than occasion. For pure facilitation a low threshold could be used, but in its access control role that would allow any traveler to board, and potentially get free passage. A conventional value for access control is for false positive rate of 1 in 10 000. Lower values can used but impostors will switch to active attack techniques to achieve a false positive. One factor is variability in false positive rates with demographics: many algorithms can give 100 times more false positives on elderly, female people from certain countries.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

Figure 6 shows the error rate tradeoff by plotting false negative identification rates against false positive identification rates at ten operating thresholds spread over four decades of FPIR, from 1 in 33 333 to nearly 1 in 3. Instead of showing the full curves, the ten-point pairs expose the increase in FNIR at low FPIR but also show the difference in error rates for men and women.

We note the following points:

1. Some algorithms give generally lower FNIR across the range of FPIR. This is simply a re-iteration that accuracy varies markedly.
2. Some algorithms give a flat error tradeoff characteristic. This is most evident for the idemia-8, deepglint-1, nec-3, paravision-5 and sensetime-4 algorithms. This is an attractive property of any biometric system because it allows very low false positive identification rates to be attained without intolerable increases in false negative identification rates. This becomes important later when we increase the enrolled population size by a factor of 100.
3. Most algorithms give FNIR below 0.03 (the green line in the plots) for a wide range of FPIR, thereby meeting the legislative mandate to be able to verify the EXIT of 97% of (in-scope) travelers.

Comparing Figure 7 with Figure 6 shows that across the four-decade range of FPIR, the FNIR values are reduced by using multiple enrollment images. The single-image enrollment represents "worst-case" of having just one prior encounter. The multi-image case is more typical.

Note that this analysis doesn't answer the technical question of whether enrolling multiple images per subject increases FPIR versus using just single image. The reason is that the thresholds for multiple enrollments are generally higher than for singles. There are exceptions – Idemia-7 for example. The question is important in situations where some travelers might have dozens of enrollment images and the algorithm response could be to attract false matches i.e. to make such enrollees lambs[15].

[15]The term lamb, a category defined in "The Biometric Zoo", refers to an enrollee who attracts more than average number of false matches.

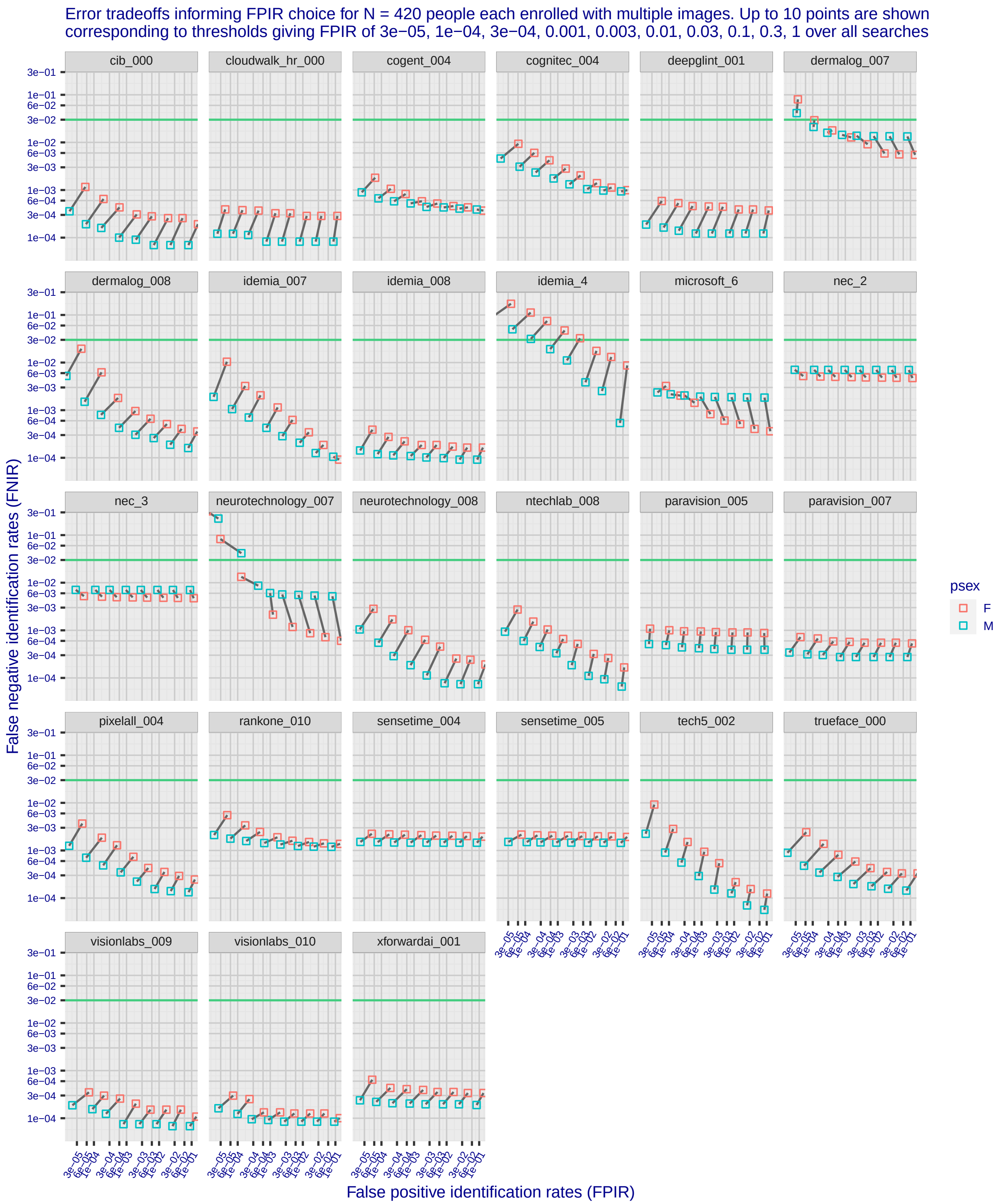


Figure 7: Error tradeoff for 26 algorithms executing 1:N searches with N = 420 people each enrolled with multiple images. The 10 point-pairs correspond to 10 possible thresholds for each algorithm. The red and blue boxes correspond to female and male travelers; their relative displacement indicates generally higher false positives and false negative rates in women. Smaller displacement indicates smaller (better) demographic differential (see NEC and Paravision-5, for example). The horizontal green line corresponds to 3% false negative goal implied by legislation in the U.S.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

3.2.4 Demographics: Differentials by sex

Vertical displacement of point pairs in Figure 6 and Figure 7 reveal broadly higher False Negative Identification Rates in women than in men. This is consistent with NIST IR 8280 using other kinds of images. The cause of this is not known. Note some algorithms, including NEC-3, Microsoft-6 and Neurotechnology-7, give the opposite behavior or fairly equitable rates.

The horizontal displacement in the figures show that all algorithms give a factor of 2 or 3 times higher false positive identification rates in women; this means that women will be mismatched against a wrong identity somewhat more often than men. This will be rare but over enough flights it will disadvantage more women than men. Algorithms from Microsoft, NEC, and Cognitec give notably smaller differentials.

Figure 8 summarizes false negative rates by sex: the difference often amounts to 1 additional false negative in women than men. Note that there are many flights with zero false negative flights for both sexes.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

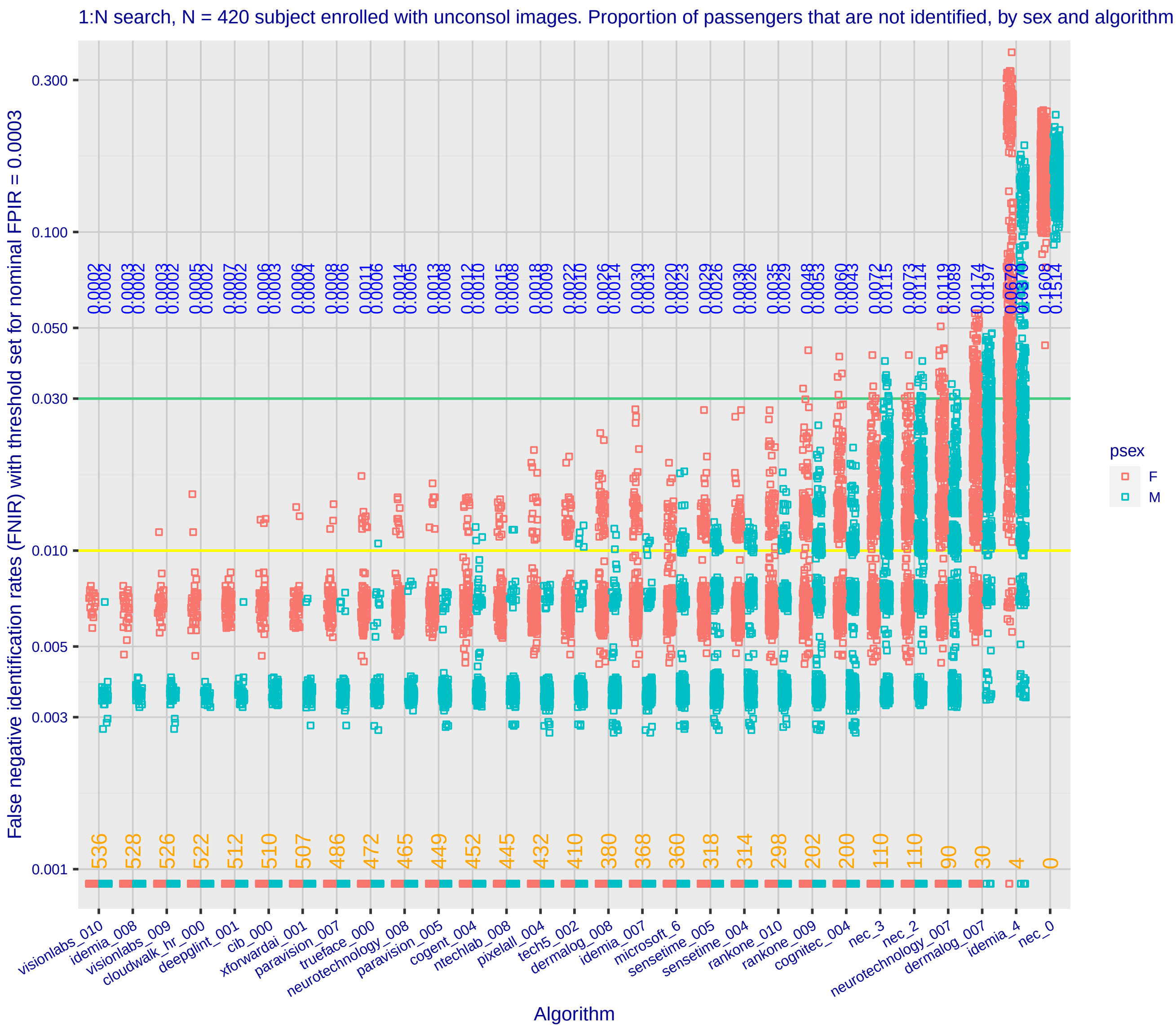


Figure 8: False negative identification rates by algorithm and sex. Each point corresponds to the boarding of N = 420 people on to one flight where each is enrolled with multiple images. The blue text is a FNIR value for that algorithm on that sex. The green line connotes a 3% FNIR (reflecting a legislative mandate). The yellow line is at 1% FNIR. The cluster of points at 0.0009 corresponds to zero errors (adjusted to plot on a log scale) - the orange text gives the number of simulated flights, out of 567, for which there are no false negative errors. The next cluster near 0.004 corresponds to 1 error out of around 210 males.

3.2.5 Demographics: False positve differentials by region

Figure 9 shows false positive identification rates by region and by sex for two algorithms from NEC and Canon. Appendix B gives analagous figures for all algorithms. We make the following comments:

1. **Magnitude:** The NEC-3 algorithms shows FPIR is quite insensitive to geography and sex with false positive identification rates estimates mostly clustered between 2x10^{-4} and 7x10^{-4} In constrast Canon's cib-000 algorithm gives FPIR estimates between 7x10^{-5} and 2x10^{-3}. As noted in NIST Interagency Report 8280 the NEC-3 algorithm is taking steps to normalize false positive rates in one-to-many searches.

2. **Sex:** It is very common across algorithms for women to give higher FPIR than men. The NEC-3 algorithm gives

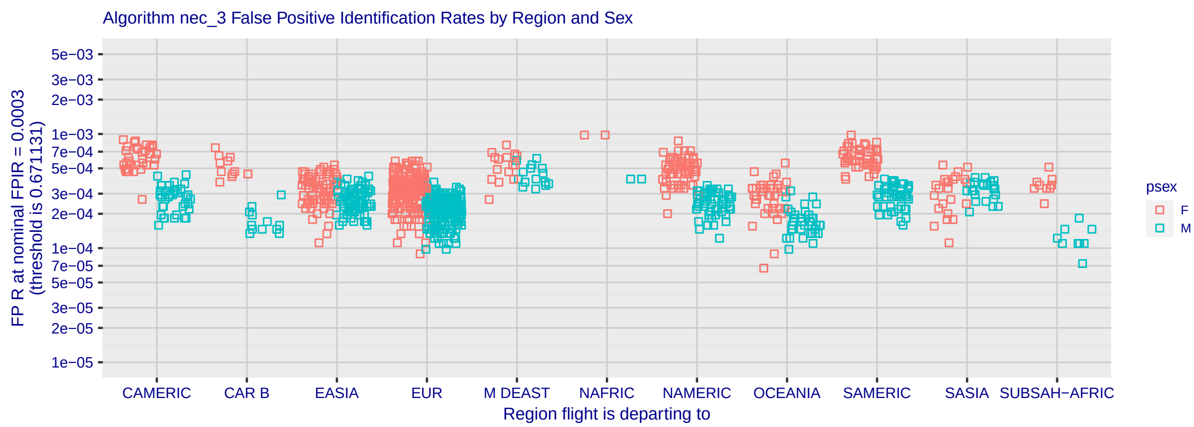
This publication is available free of charge from: https://doi.org/10.6028/NIST IR 8381

broadly the smallest differential in FPIR value – cf. the y-axis in the next Figure.

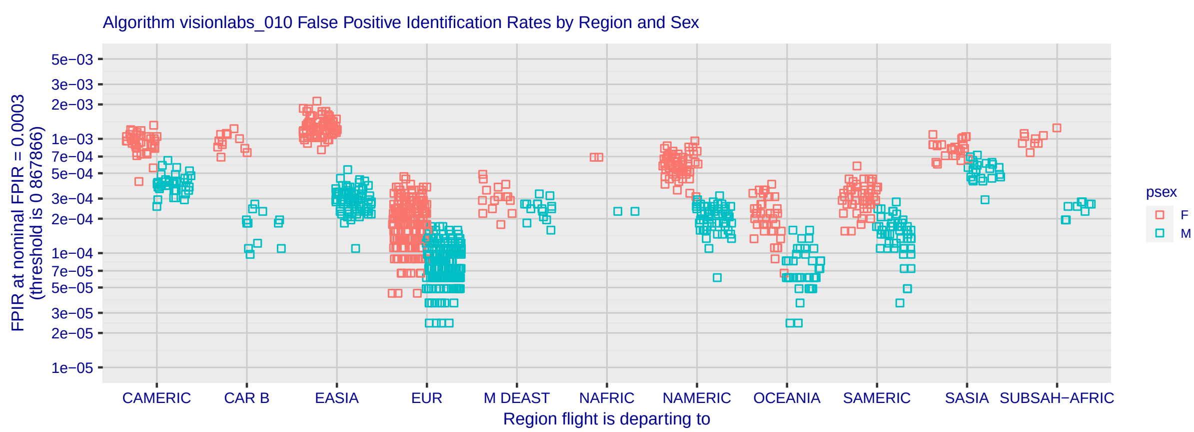
3. **Region:** False positive identification rates are commonly an order of magnitude higher in Asian women than in European men. For Canon's cib-0 algorithms there is a factor of 30 variation.

As always, with the observation of a demographic differential, the question is "what is the impact"? The overall target FPIR was 0.0003, achieved by setting an algorithm specific target. The worst upside departure from that is Canon's cib-1 algorithm (see Appendix B) which gives FPIR for Asian women near 0.003. This FPIR, 1 in 333, is still low but implies between one and two false positives per flight boarding – these would likely manifest as an in-gallery false match described in section 1.5. This may be an acceptable cost, but does constitute a disadvantage for Asian women attempting to record their departure from the United States.

The error tradeoff characteristics of figure 11 are, for some algorithms, quite flat implying that even lower false positive identification rates could be targetted (by increasing the threshold) without great adverse implications for false negative identification rates.



(a) NEC-3



(b) VisionLabs-10

Figure 9: For two algorithms, each point shows a false positive identification rate estimated by running c. 120 000 searches against that flight's gallery. Red and blue connote male and female enrollees. Analogous figures for all algorithms appear in Appendix B.

4 Scaling One-To-Many Authentication to Larger Populations

4.1 Motivation

Thus far we have discussed the use of 1:N face recognition for recording the exit of travelers while boarding an aircraft. A PCA can appropriately limit enrollment in FR galleries to just the population expected on the flight. This data minimization reduces mis-match possibilities. However, the travel industry has articulated a vision for paperless travel. In its simplest form, this starts when a traveler authenticates to an authoritative travel document (passport) using a one-to-one biometric verification of a live photo and then proceeds through an airport's "touchpoints" such as the TSA line and airline lounges and aircraft boarding, without presenting a boarding pass. Instead, the traveler engages a camera which submits a photo as a query into a database of individuals expected and authorized to proceed. For example, such a system could be fielded at a security checkpoint. In such cases many more people would need to be enrolled into the face recognition engine than at a departure gate – for example, all people expected in the airport during a time window extending from a few hours before their respective flights to the time of expected or actual departure. The number of individuals could readily extend into the tens of thousands, and more if airside locations would additionally recognize inbound passengers (e.g. buying in duty-free shops).

4.2 Background

The dependence of recognition accuracy on enrolled population size is well known. Qualitatively, as enrolled population grows any given search has a greater possibility of a false match. Such outcomes can occur for two types of traveler.

1. An illegitimate traveler – someone who is not expected in the airport – makes a presentation to the camera in attempt to pass the checkpoint. This succeeds if the traveler matches any enrolled identity with a comparison score above threshold.
2. A legitimate traveler – someone who is expected at the touchpoint – presents to the camera but matches an identity other than self. This may be inconsequential at a TSA line, but would be consequential in a hypothetical duty-free store application of this approach should the biometric result allow purchase without further authentication.

The rate at which false positives occur is the false positive identification rate (FPIR). In a biometric test, FPIR is estimated by conducting non-mated searches into an enrolled population. FPIR is stated as the number of searches resulting in a false positive divided by the number of non-mated searches. How does FPIR scale with the number of enrolled identities? There are two classes of face search algorithms: Class A is those that implement a 1:N search as N 1:1 comparisons followed by a sort operation, and Class B is comprised of everything else including those that implement some more complex search strategy.

- **Class A** algorithms are expected to give a FPIR increases with the number of enrolled identities. It may increase further also if those identities are enrolled with several images each. Given a system in which N people are enrolled, with one image each, a standard binomial model gives

$$\mathrm{FPIR}(N,T) = 1 - (1 - \mathrm{FMR}(T))^N \tag{3}$$

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

where the system owner sets the threshold T, and has an estimate of FMR(T), the false match rate in purely one-to-one comparisons. For small FMR, this approximates to

$$\mathrm{FPIR}(N,T) = N\mathrm{FMR}(T) \tag{4}$$

implying that the one-to-many false positive hazard grows linearly with N.

- **Class B** one-to-many algorithms are those that do not implement 1:N search using N 1:1 comparisons - these can include fast-search algorithms (using trees, indexes etc) and those that normalize scores across some or all of the gallery entries. These algorithms may not exhibit the (near) linear dependence of equations 3 and 4. This can occur for other reasons also. Some algorithms adjust comparison scores to the database size such that FPIR becomes approximately independent of N. The NEC-3 and Idemia algorithms exhibit such behavior (see FRVT Part 2 and its report cards). This relieves the system owner of the need to configure thresholds for the given population size. A system owner might consult vendor documentation, or consult NIST's FRVT Part 2 report which documents the dependence of FPIR on N and T.

4.3 Simulation of Large-N Accuracy

4.3.1 Experimental design

We repeated the EXIT simulation given previously but instead of enrolling N = 420 individuals with one ENTRY image, we mixed in a further 41580 such images from a disjoint population selected without regard to demographics. The result is a set of 567 galleries, each with N = 42 000 individuals. This population size is somewhat larger relative to the number of international passengers appearing daily in large U.S. airports in 2019.

4.3.2 Results

Figure 10 shows the number of false negatives expected when using the algorithm named on the horizontal axis to search three different kinds of galleries. The first enrolls 420 people with a single PCA ENTRY image; the second enrolls those same people with all prior PCA ENTRY encounters; the last enrolls 42 000 people with a single image. The kind of gallery is encoded by the shape. The vertical position of each point is the mean (over 567 regional galleries) of the number of false negatives when 420 test subjects are searched. The color of each point encodes the fraction of all 567 trials that give three or fewer false negatives.

We make the following observations:

1. The number of false negatives is higher with N = 42 000 than with N = 420, as expected.

2. Some algorithms nevertheless give only modest increases in the number of false negatives. In the most accurate case, the mean number of passengers being rejected would be below 1% (4/420), and more than 75% of flights (trials) would have three or fewer false negatives out of the 420 people making attempts.

3. Other algorithms give substantially higher false negative rates – the graph shows FNIR approaching about 8% (34/420) for legacy algorithms.

4. Note that this analysis does not consider variance around the point estimates, nor sex or regional differences.

4.3.3 Discussion

Large enrolled populations require algorithms to be configured to operate at lower false match rates – following Equation 4 an increase in N from 420 to 42 000 will necessitate a 100-fold FMR reduction to maintain constant FPIR. This puts a premium on algorithms that maintain relatively low FNIR at lower FPIR.

By inspecting Figure 11, for N = 42 000, all algorithms except some from Cloudwalk, Deepglint, Idemia, NEC, Paravision, Sensetime, Visionlabs and X-ForwardAI cannot maintain FNIR below 0.03 so, depending on FPIR, would not be meeting a legislative mandate for FNIR < 0.03 and TPIR > 0.97.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

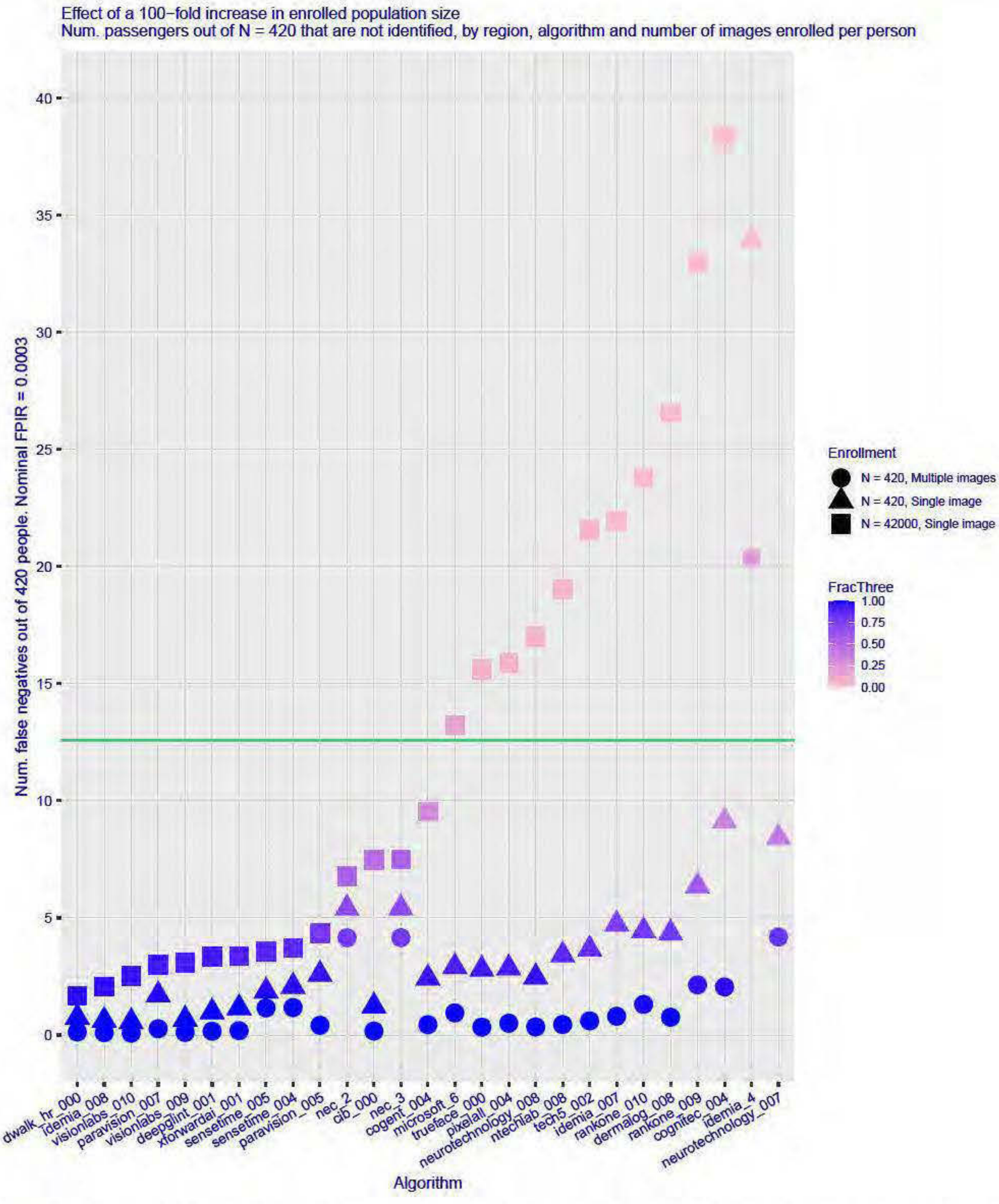


Figure 10: Comparing the number of false negatives expected when N = 420 people are searched against galleries containing imagery from enrolled populations of size of N = 420 and N = 42 000. The threshold is fixed in all cases to produce a false positive identification error rate of 1 in 3 333 (FPIR = 0.0003). The threshold value for each algorithm will usually be higher for the larger gallery to maintain the same false positive likelihood. The horizontal green line corresponds to a 3% false negative rate.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

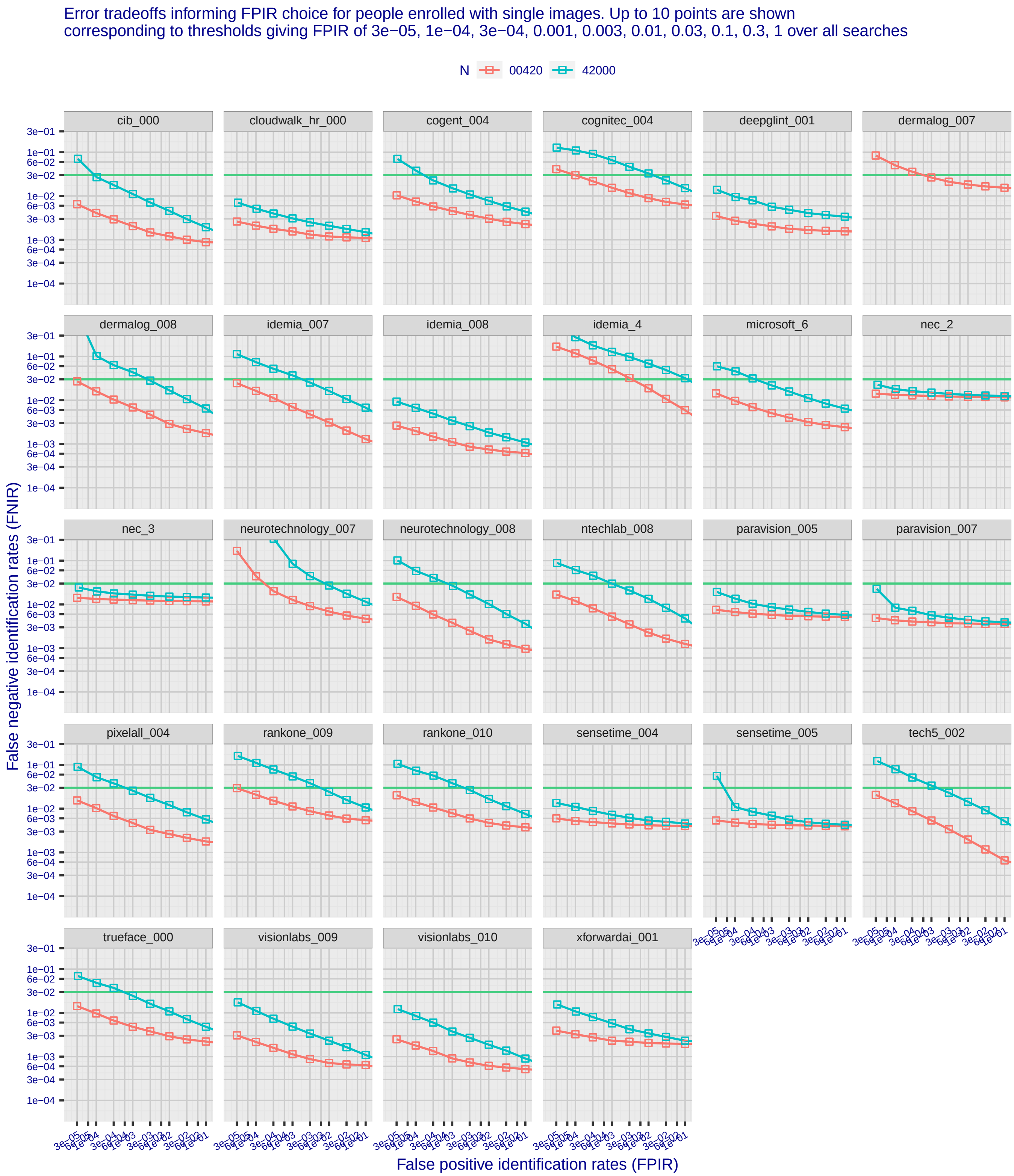


Figure 11: Error tradeoff characteristics for twelve algorithms conducting identical sets of searches into galleries of size N = 420 and N = 42 000. The horizontal line corresponds to a 3% false negative identification rate. The left side of each panel is relevant to the more "lights-out" use of FR in positive access control and EXIT facilitation; the right side of each panel corresponds to high false positive identification rates for investigative uses of FR where humans review candidate lists. A flat profile confers the advantage of being able to run at lower FPIR without much elevation in FNIR.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

5 Factors That Render Accuracy Estimates Approximate

The result in this report do not constitute an answer to the questions "how well does a particular TVS work", "does TVS satisfy a 97% legislative verification mandate" or "what is the accuracy of a PCA's EXIT solution". Why? Because the questions are different and because the tests we have reported here, while extensive, depart from the intended and desirable tests as follows. For each factor discussed, we note in blue the expected effect on accuracy.

1. **Airline re-direction of passengers.** During the EXIT pilot, airlines diverted some customers toward the legacy paper-based boarding process. This was particularly true when boarding was proceeding slowly or when cameras of the network to TVS were malfunctioning. This is not expected to bias accuracy in an offline test either way, but would lead to complication in using TVS logs to measure accuracy.

   The population so diverted was sometimes not random – very tall or short travelers, and those with children, would be directed from the FR line. To the extent that this occurs, and to the extent that the NIST EXIT collection is not itself affected by this, our estimates of accuracy may be too high.

2. **Algorithms:** NIST does not have access to the actual algorithms deployed in TVS systems. Instead this study uses prototypes submitted to the one-to-many search track of the FRVT. These prototypes are identified using a name and a number. For example, "NEC-3" is from the NEC Corporation, and the three is simply a sequence number of algorithms sent to NIST. NIST is unable to confirm whether any prototype in FRVT has ever been deployed. Indeed a developer may make decisions on whether to productize a prototype on the basis of FRVT-derived technical information. In any case, a developer will maintain their own versioning designations. NIST is not provided with copies of operational algorithms.

3. **Active development**: Given persistent improvements in accuracy, as documented in FRVT, it is incumbent on end-users to instantiate a "technology-refresh" procedure so as to realize accuracy gains. Note that results in this report for 2018-era algorithms are likely out-of-date. Thus, for any given developer, it is likely that higher accuracy is available than is estimated here.

4. **Algorithm post-processing:** Accuracy will change if any software is used to post-process candidate lists produced by the algorithm. Conventionally the face recognition algorithm issues a candidate identity and a similarity score, which is compared to a system-wide threshold. If post-processing is used to re-score or re-rank then its effect on both false positive and false negative identification rates should be measured by comparing with that available from the raw candidate lists alone.

5. **Image data:** International travel has long been predicated on presentation of a passport. With e-Passports it is common for passport images to be retrieved and used for 1:1 verification of the the traveler. If on ENTRY those images are retained by the PCA , they can be used in downstream EXIT face recognition processes. The same applies to visa portraits collected as part of a visa application.

   (a) We did not have passport or passport-equivalant images for use in this study. These include visa images of various travelers and passport photos of the citizens persons. Instead we used airport arrivals hall photos with reduced quality. To the extent that a TVS makes use of high-quality passport and visa images, the accuracy values reported here are likely to be worse than for a system for which such images are available.

   (b) In this study we used an extract of a much larger corpus provided to NIST in May 2019. These images were anonymized and accompanied by limited metadata. The set included images labelled *exit*, and *ENTRY*. The former were collected in airport departures. A few of the exit images appear to have been collected from persons in a vehicle, as could occur at a land-border. This factor will tend cause our accuracy estimates to differ from those of an operational TVS.

(c) Our exit images were collected in 2018 and the first four months of 2019. We assume that subsequent cameras, and their refined deployment by airlines, will yield improved images today compared to those used in this study, so we would expect improved accuracy over that noted here.

(d) Moreover, our exit images are not accompanied by camera make and model information, nor flight manifest information. It was therefore not possible for NIST to exactly reconstruct "a flight" – instead we pooled all exit images as probes searched against each gallery. Our search set therefore pools exit images from quite different cameras and locations (airports). We are therefore unable to compare cameras and collection sites. From observations made during site visits, we note markedly different approaches to the quality-speed tradeoff. We expect therefore that our accuracy estimates have reduced variance compared to that from a TVS.

(e) If exit images are retained, even for a short period, they may be useful in offline "after-hours" accuracy estimation. For example, images from one flight could be used to make non-mated searches into a gallery of another flight, so as to estimate FPIR.

(f) Our galleries were constructed to hold people from one travel region as inferred from the nation that issued their travel document. This means that the galleries in this document will contain people who never flew together.

6. **Homogenous galleries.** Our practice of making galleries from people holding travel documents from countries in the same region of the world probably means that false positive rates are higher than if the galleries had been composed of a more mixed population. This practice would tend to depress our accuracy relative to those in TVS.

7. **Presence of images active attack.** It is possible that some of the captured images are from a presentation attack that went undetected, for example using a face mask. The occurrence of this is considered to be very small. Note that since we didn't have passport images, we do not expect the dataset to contain morphed images. While this is increasing possibility operationally, it can be averted by live, trusted capture of images as in a primary passport control lane.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

Appendices

Appendix A Figures summarizing false negatives for each algorithm

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

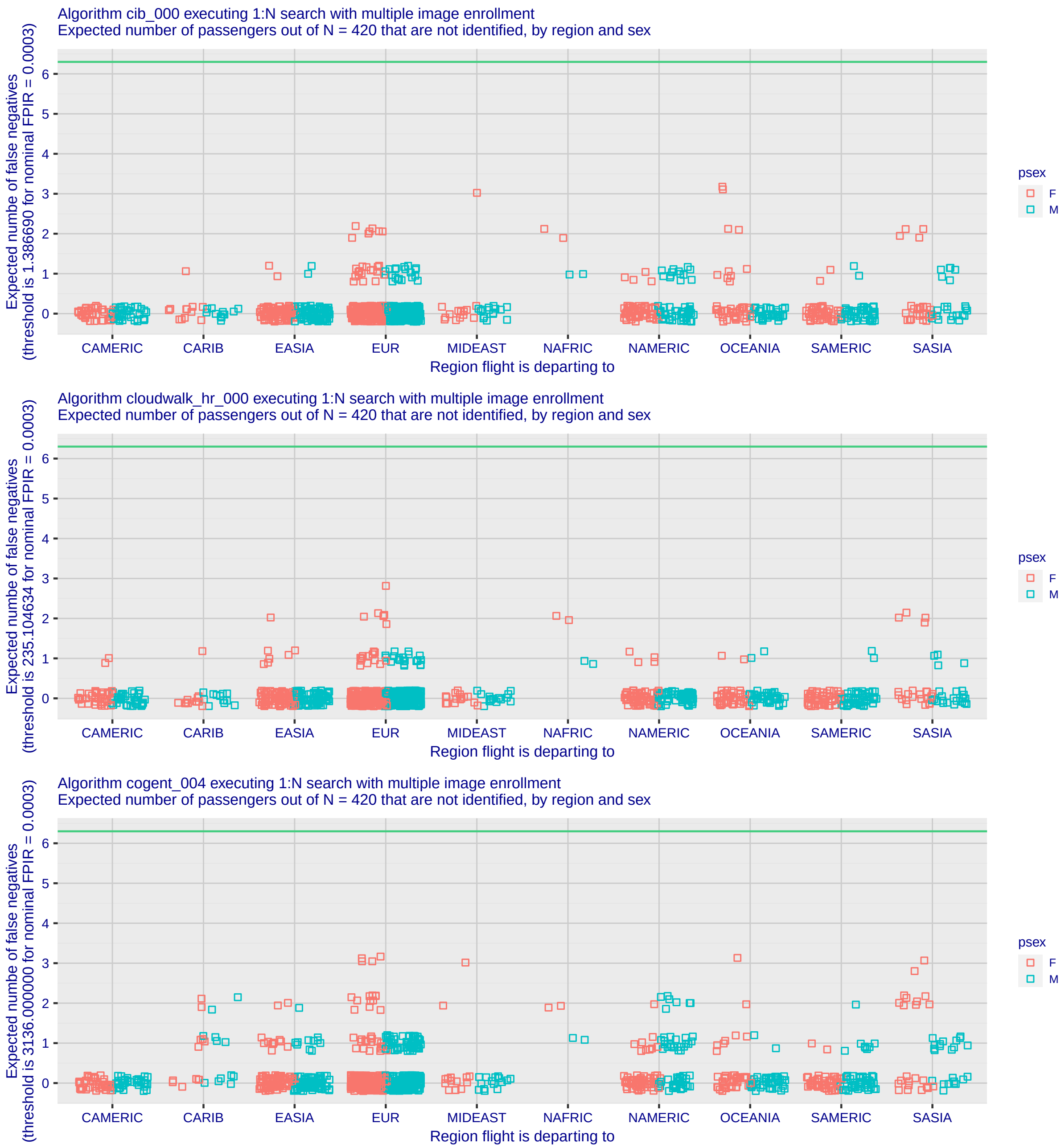


Figure 12: For the eleven regions and two sexes, each point give the expected number of false negatives for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The numbers are stated by scaling measured numbers of false negatives to 210 per sex. The points' positions are jittered horizontally and vertically to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have. The number of individuals in the gallery is exactly 420.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

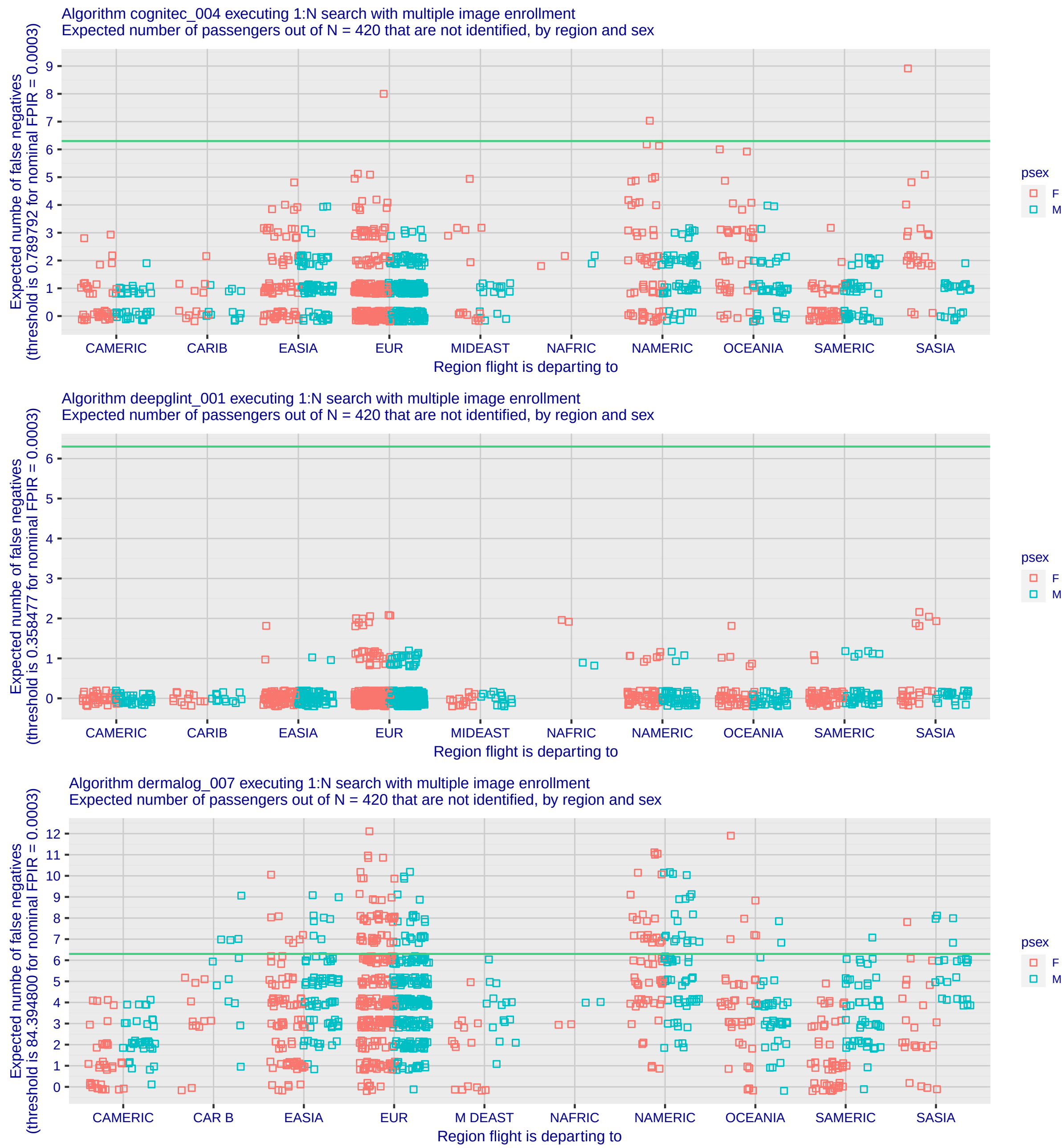


Figure 13: For the eleven regions and two sexes, each point give the expected number of false negatives for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The numbers are stated by scaling measured numbers of false negatives to 210 per sex. The points' positions are jittered horizontally and vertically to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have. The number of individuals in the gallery is exactly 420.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

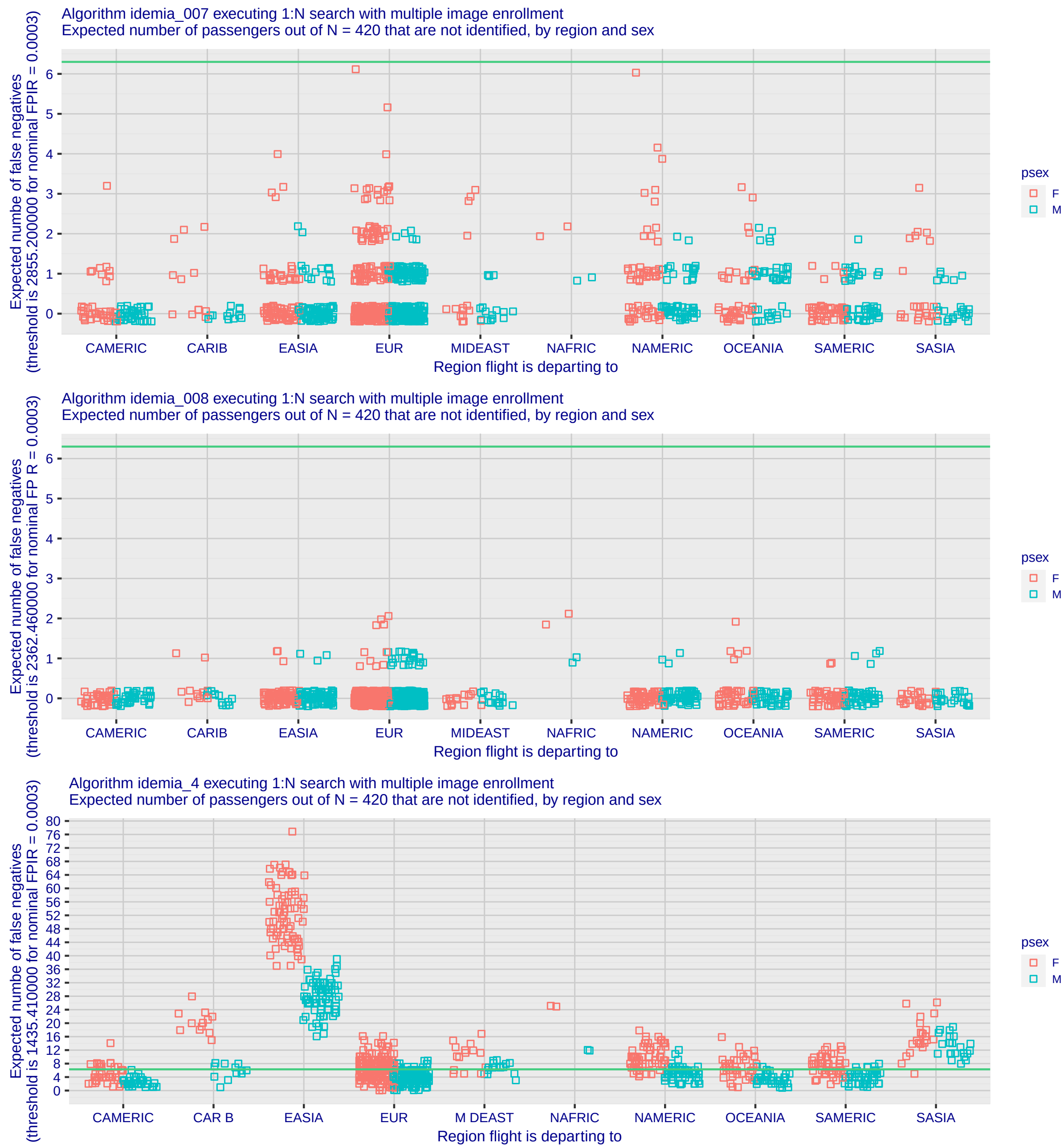


Figure 14: For the eleven regions and two sexes, each point give the expected number of false negatives for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The numbers are stated by scaling measured numbers of false negatives to 210 per sex. The points' positions are jittered horizontally and vertically to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have. The number of individuals in the gallery is exactly 420.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

Figure 15: For the eleven regions and two sexes, each point give the expected number of false negatives for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The numbers are stated by scaling measured numbers of false negatives to 210 per sex. The points' positions are jittered horizontally and vertically to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have. The number of individuals in the gallery is exactly 420.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

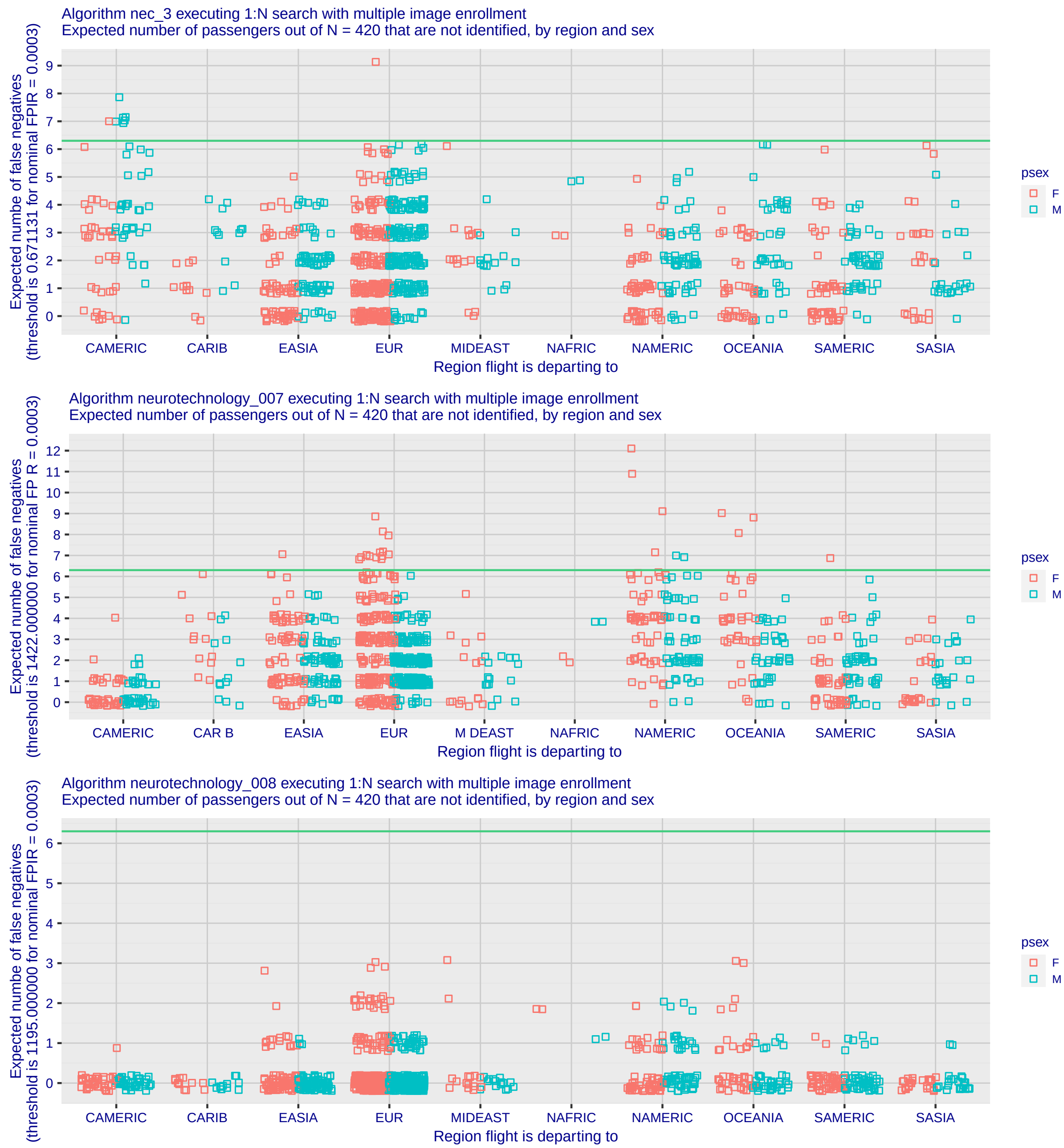


Figure 16: For the eleven regions and two sexes, each point give the expected number of false negatives for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The numbers are stated by scaling measured numbers of false negatives to 210 per sex. The points' positions are jittered horizontally and vertically to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have. The number of individuals in the gallery is exactly 420.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

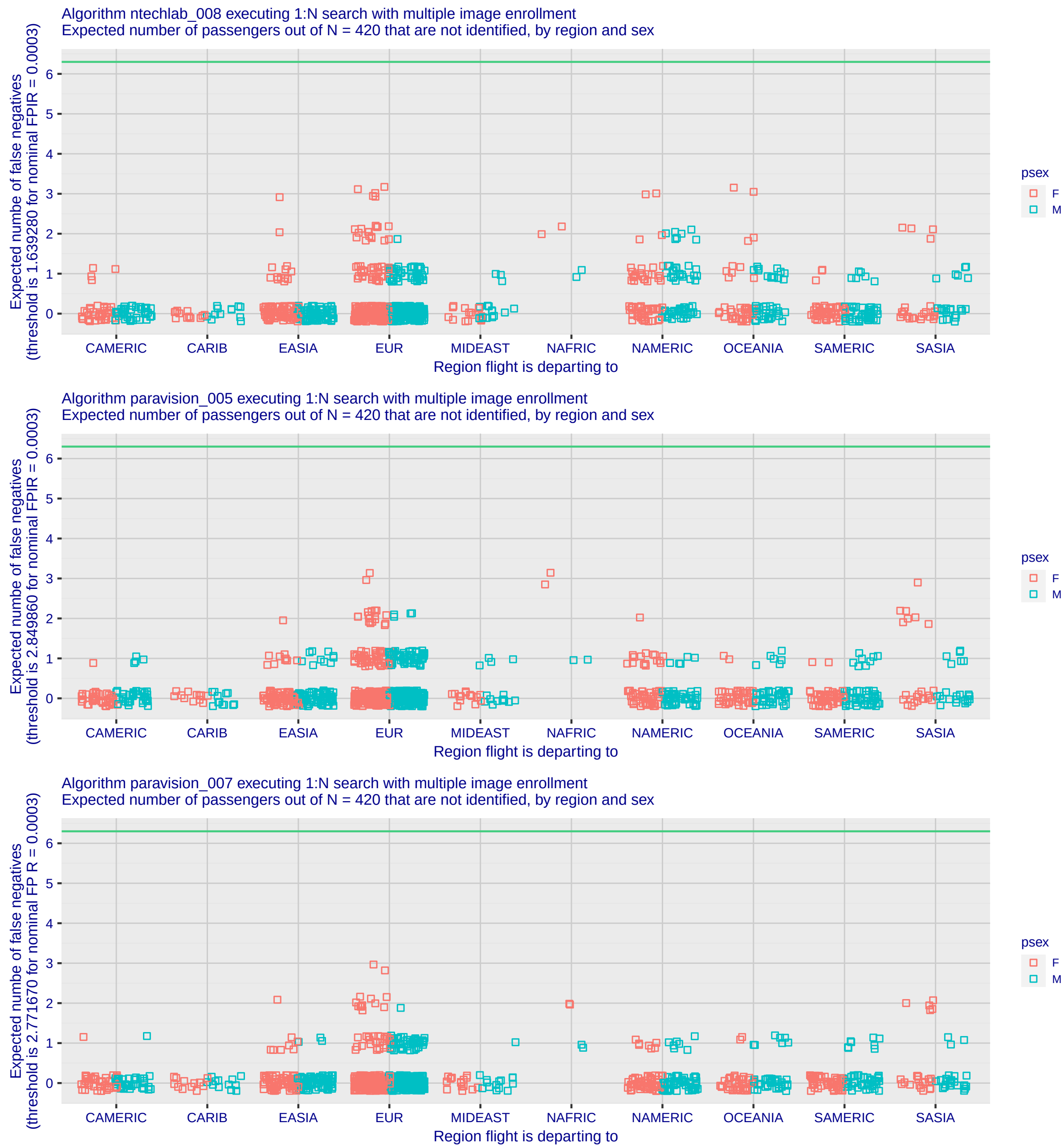


Figure 17: For the eleven regions and two sexes, each point give the expected number of false negatives for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The numbers are stated by scaling measured numbers of false negatives to 210 per sex. The points' positions are jittered horizontally and vertically to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have. The number of individuals in the gallery is exactly 420.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

Figure 18: For the eleven regions and two sexes, each point give the expected number of false negatives for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The numbers are stated by scaling measured numbers of false negatives to 210 per sex. The points' positions are jittered horizontally and vertically to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have. The number of individuals in the gallery is exactly 420.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

Figure 19: For the eleven regions and two sexes, each point give the expected number of false negatives for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The numbers are stated by scaling measured numbers of false negatives to 210 per sex. The points' positions are jittered horizontally and vertically to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have. The number of individuals in the gallery is exactly 420.

This publication is available free of charge from: https://doi org/10 6028/NIST IR 8381

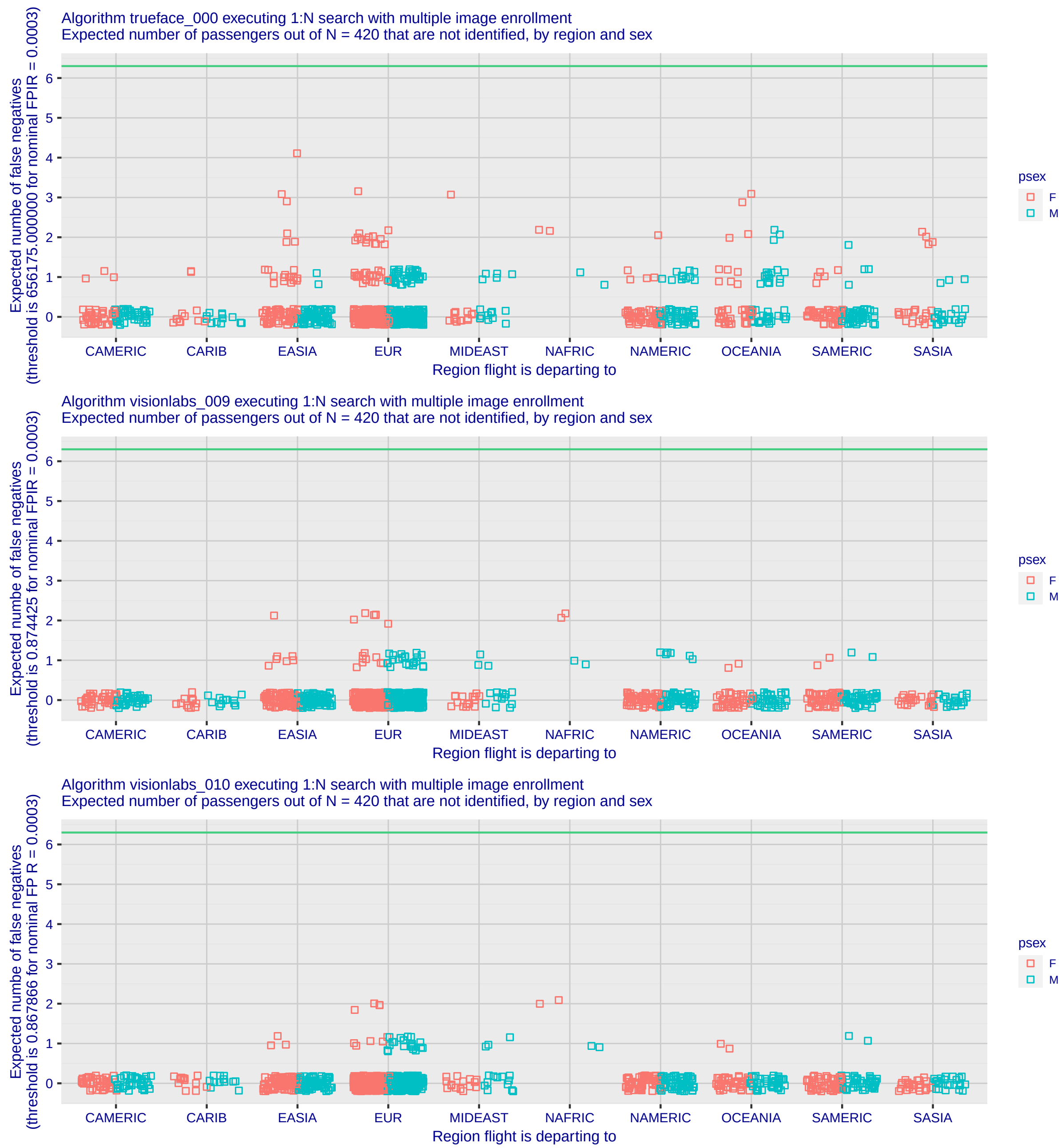


Figure 20: For the eleven regions and two sexes, each point give the expected number of false negatives for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The numbers are stated by scaling measured numbers of false negatives to 210 per sex. The points' positions are jittered horizontally and vertically to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have. The number of individuals in the gallery is exactly 420.

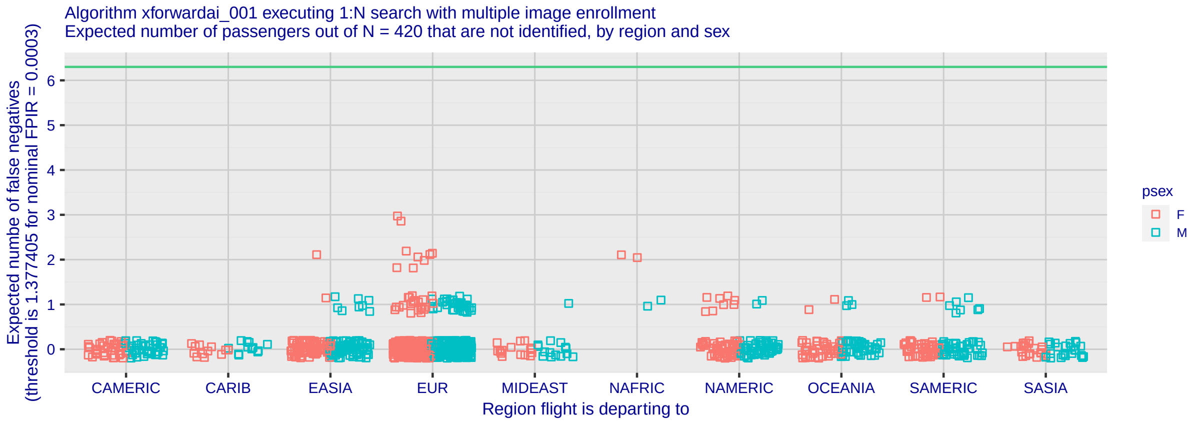
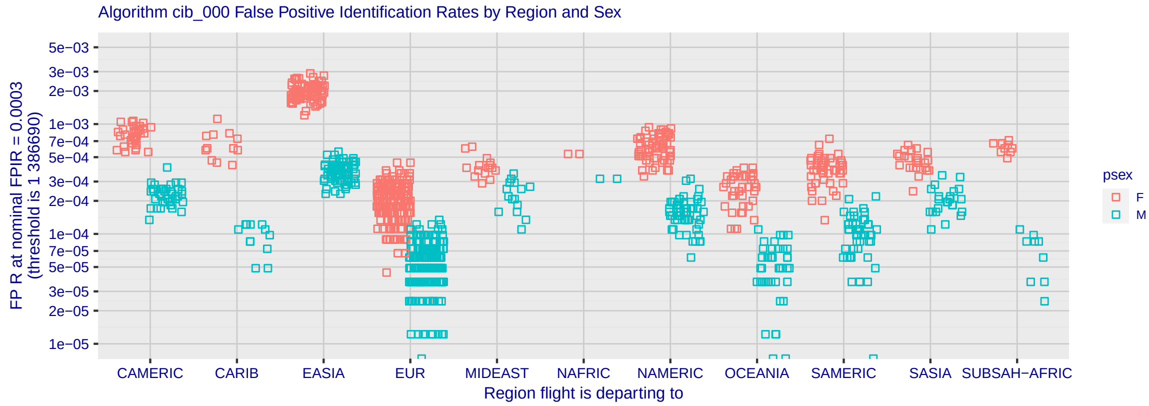


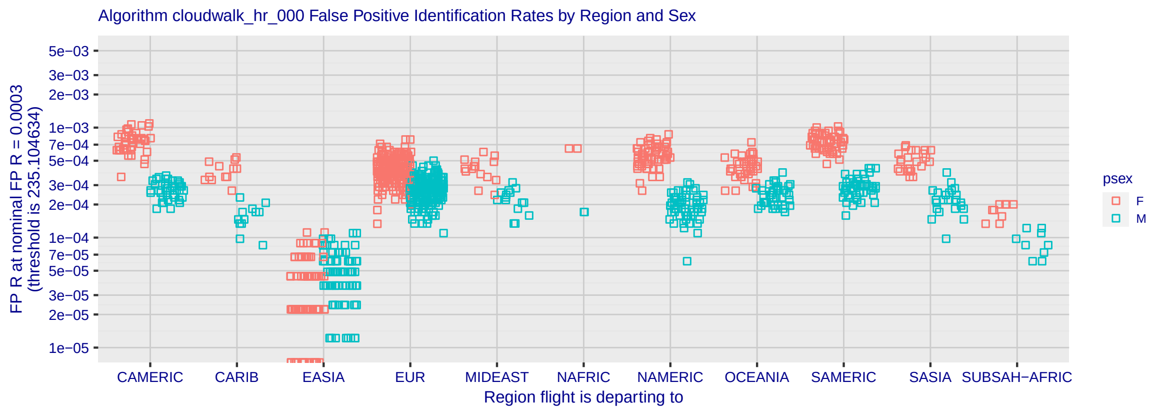
Figure 21: For the eleven regions and two sexes, each point give the expected number of false negatives for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The numbers are stated by scaling measured numbers of false negatives to 210 per sex. The points' positions are jittered horizontally and vertically to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have. The number of individuals in the gallery is exactly 420.

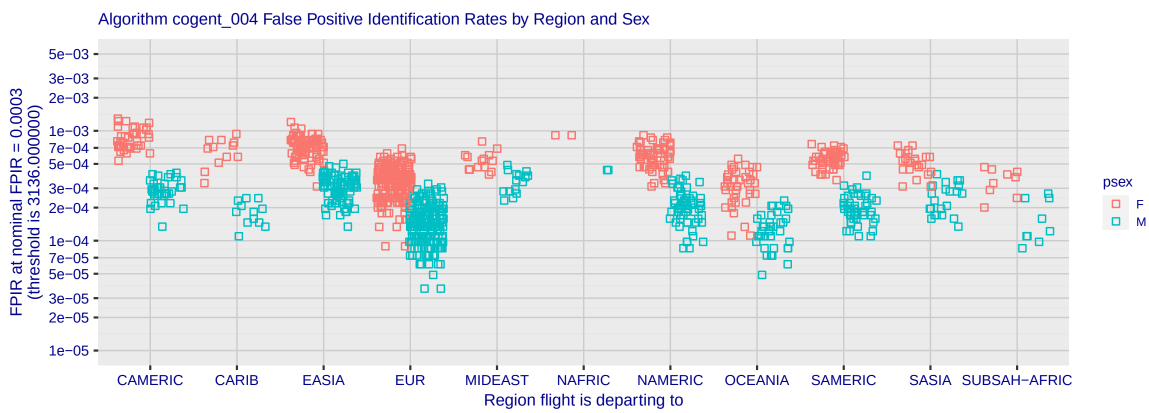
Appendix B Figures summarizing false positive identification rate for each algorithm

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381







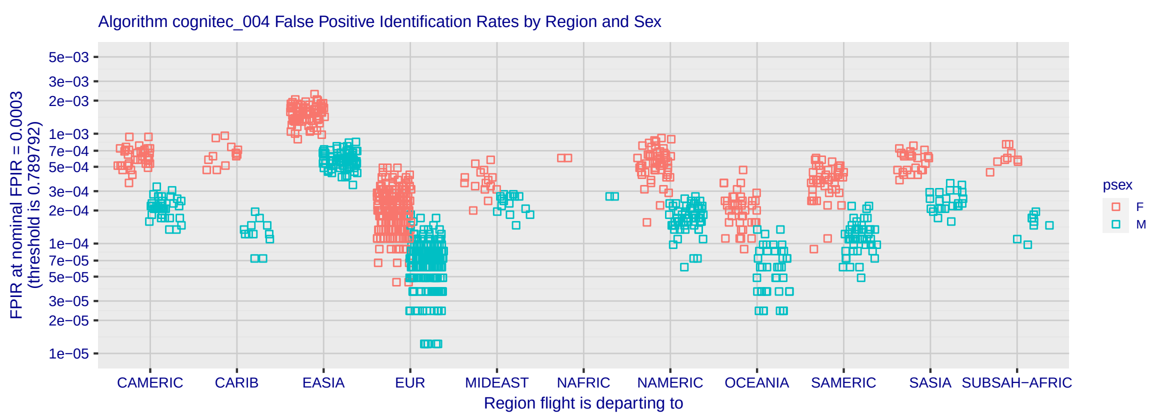


Figure 22: For the eleven regions and two sexes, each point give the false positive identification rarte for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The points' positions are jittered horizontally to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have The number of individuals in the gallery is exactly 420.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

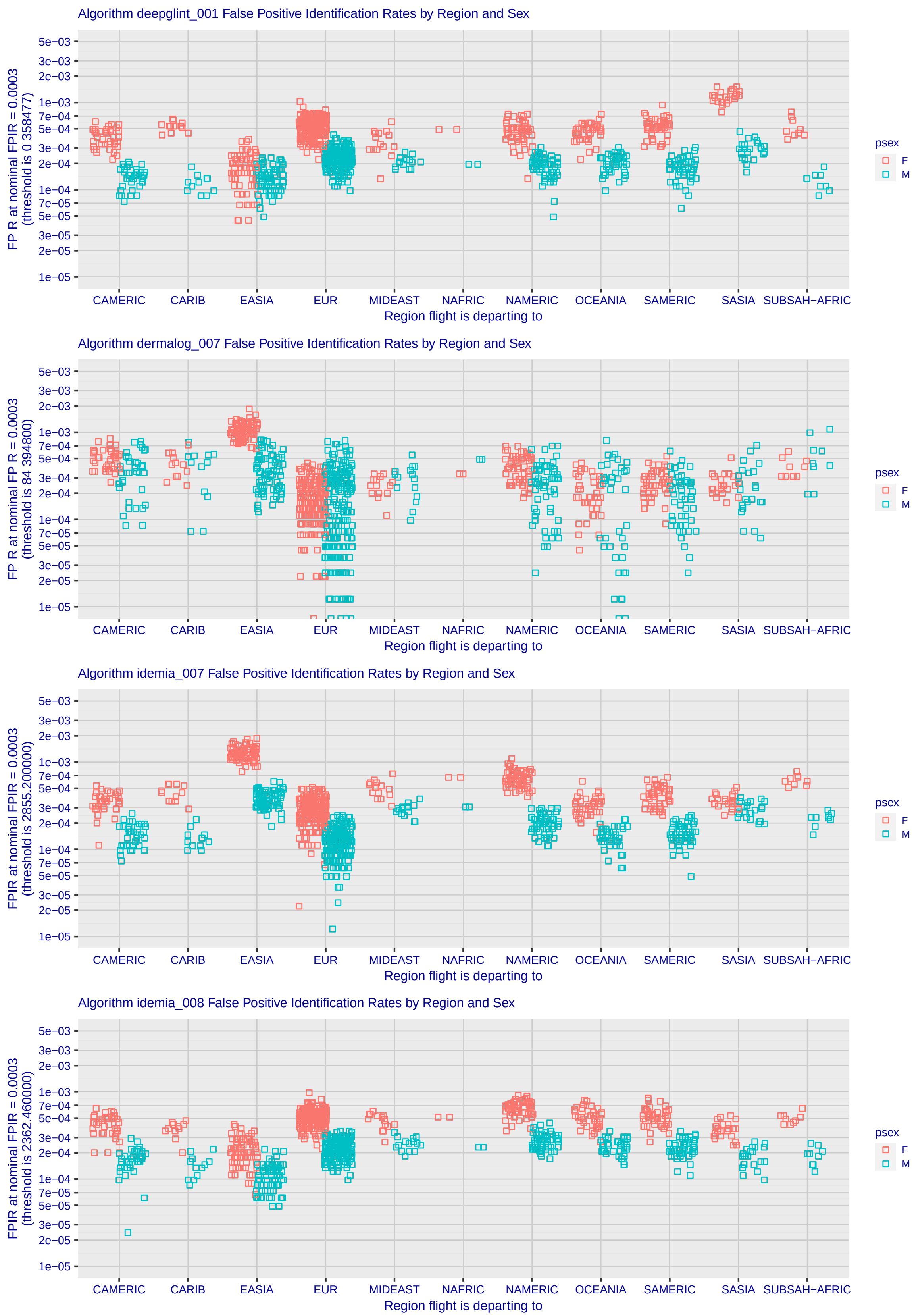


Figure 23: For the eleven regions and two sexes, each point give the false positive identification rarte for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The points' positions are jittered horizontally to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have The number of individuals in the gallery is exactly 420.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

Figure 24: For the eleven regions and two sexes, each point give the false positive identification rarte for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The points' positions are jittered horizontally to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have The number of individuals in the gallery is exactly 420.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

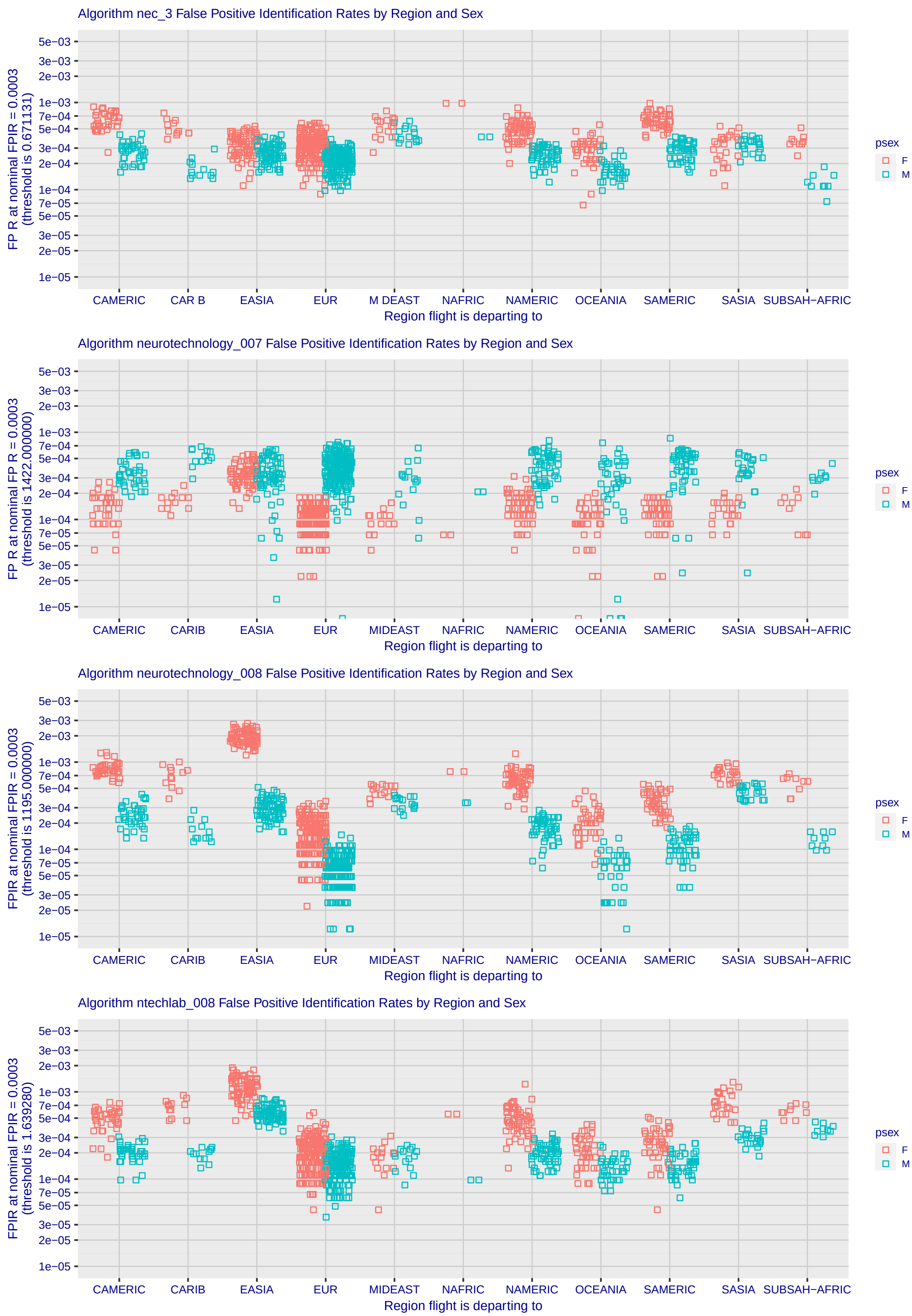


Figure 25: For the eleven regions and two sexes, each point give the false positive identification rarte for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The points' positions are jittered horizontally to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have The number of individuals in the gallery is exactly 420.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

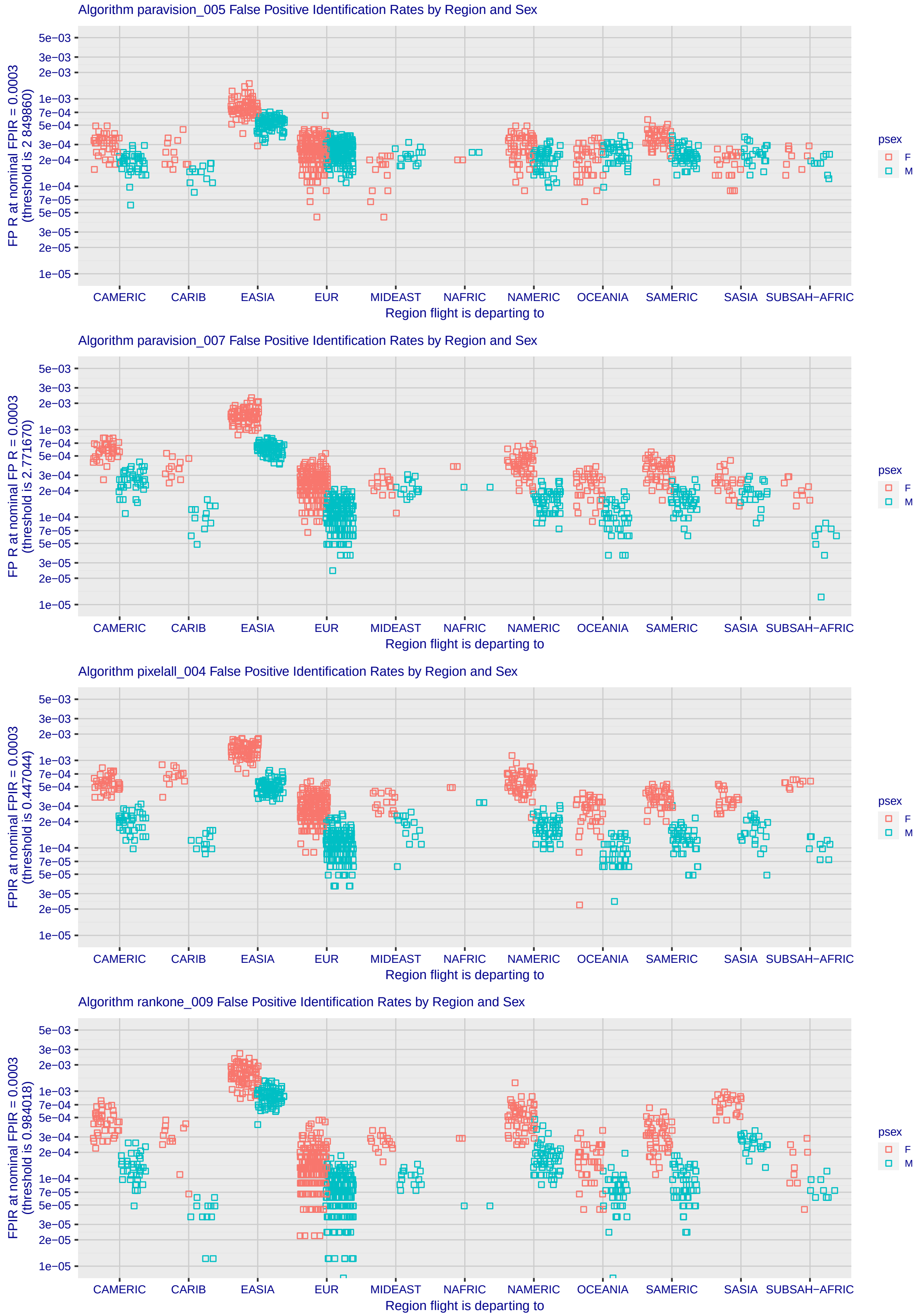


Figure 26: For the eleven regions and two sexes, each point give the false positive identification rarte for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The points' positions are jittered horizontally to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have The number of individuals in the gallery is exactly 420.

This publication is available free of charge from: https://doi org/10 6028/NIST IR 8381

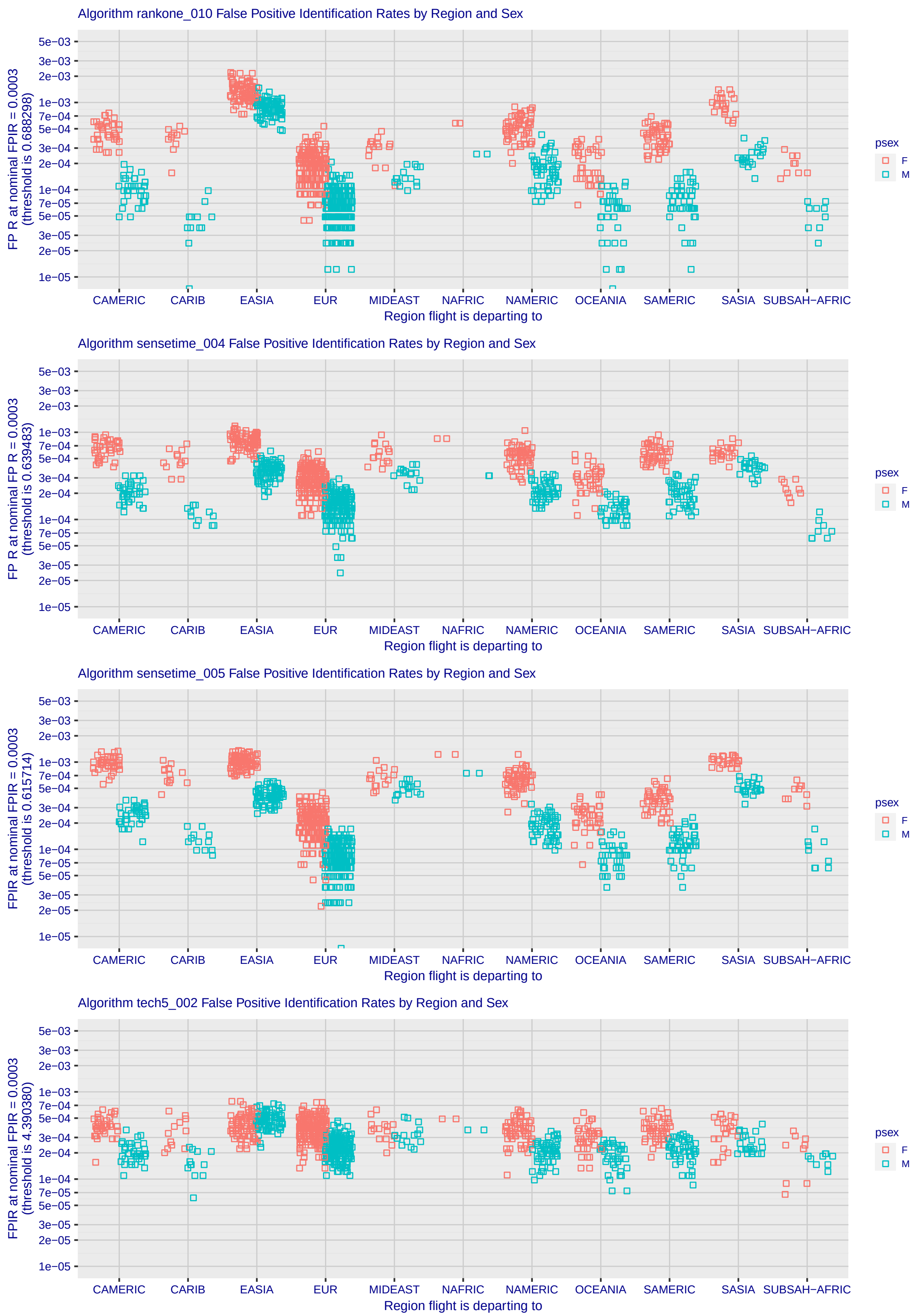


Figure 27: For the eleven regions and two sexes, each point give the false positive identification rarte for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The points' positions are jittered horizontally to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have The number of individuals in the gallery is exactly 420.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381

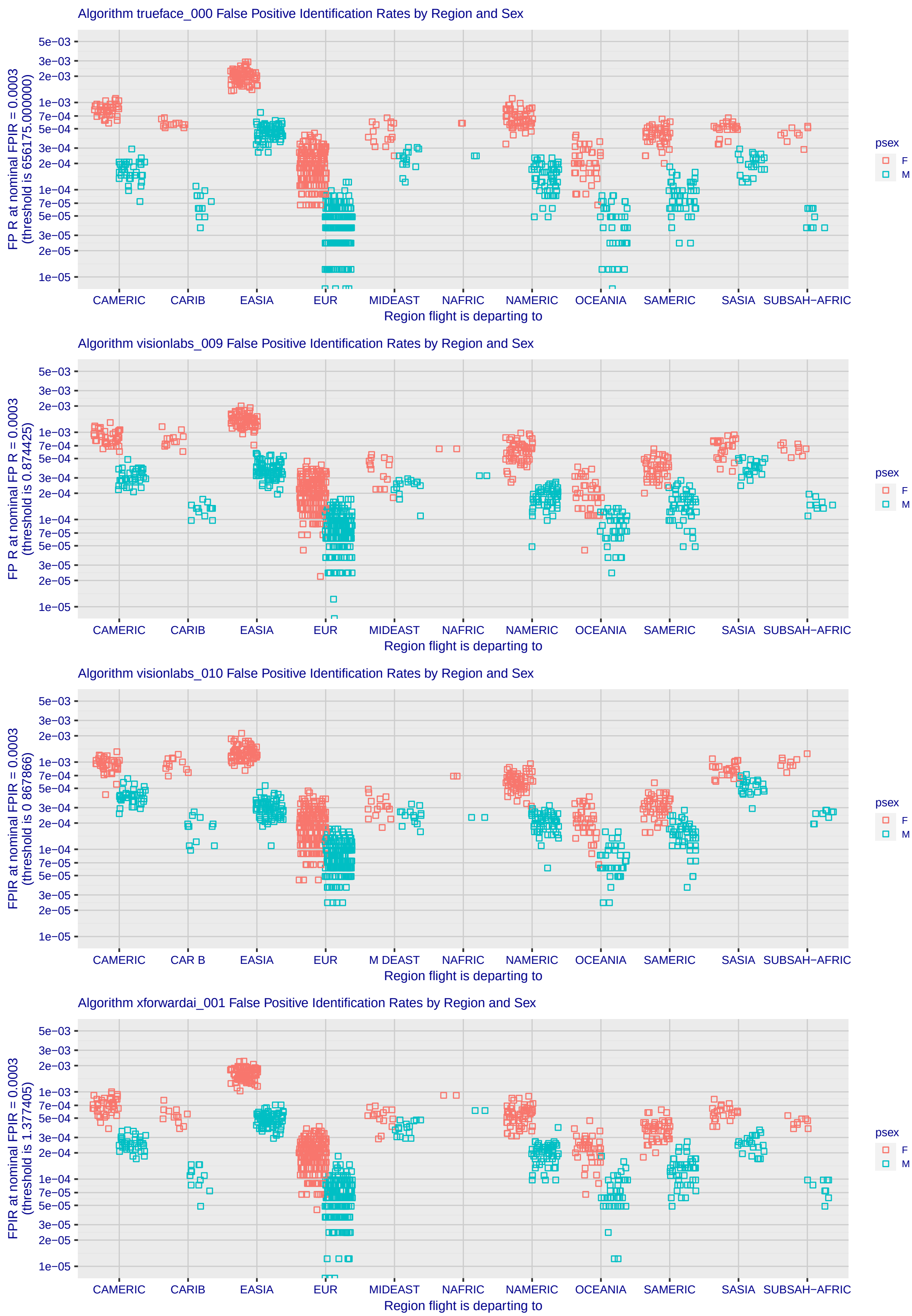


Figure 28: For the eleven regions and two sexes, each point give the false positive identification rarte for a simuliated flight in which 420 passengers, 210 men and 210 women, attempt boarding after being enrolled with multiple images each. The points' positions are jittered horizontally to mitigate over-plotting invisibility. There are many more flights to Europe, particularly, and East Asia simply because of their representation in the EXIT image corpus we have The number of individuals in the gallery is exactly 420.

References

[1] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test (frvt) part 2: Identification. Interagency Report 8271, National Institute of Standards and Technology, Home: https://pages.nist.gov/frvt/html/frvt1N.html, September 2019. https://doi.org/10.6028/NIST.IR.8271.

[2] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test (frvt) part 3: Demographic effects. Interagency Report 8280, National Institute of Standards and Technology, Home: https://pages.nist.gov/frvt/html/frvt11.html, December 2019. https://doi.org/10.6028/NIST.IR.8280.

This publication is available free of charge from: https://doi.org/10.6028/NIST.IR.8381