IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, September 2026
Impact Factor: 8.2

Role of Performance Evaluation Metrics and Benchmark Datasets for Fingerprint-Based Biometric Authentication

Neetu[1] and Dr. Sankalp Mehandiratta[2]
[1]Research Scholar, Department of Computer Application
[2]Assistant Professor, Department of Computer Application
Sunrise University, Alwar, Rajasthan

Abstract : *Fingerprint recognition is widely used for personal verification and identification, yet reported accuracy varies substantially with sensor characteristics, image quality, population, acquisition conditions, matching algorithms, and evaluation protocol. This paper examines the role of performance metrics and benchmark datasets in producing reliable, reproducible, and interpretable assessments of fingerprint-based biometric authentication. It reviews threshold-based measuresincluding false match rate, false non-match rate, equal error rate, receiver operating characteristic and detection error trade-off curves as well as identification measures, failure rates, image-quality indicators, and operational throughput. It also discusses benchmark resources such as the Fingerprint Verification Competition databases, NIST Special Database 4, and latent-print collections.*

Keywords: Fingerprint authentication, biometric evaluation, FMR, FNMR, EER, benchmark datasets, FVC, NIST, reproducibility.

1. INTRODUCTION

Biometric authentication verifies identity through physiological or behavioral characteristics. Fingerprints remain a prominent modality because friction-ridge patterns are distinctive, capture devices are widely available, and matching can be performed in compact systems. However, a fingerprint matcher is not evaluated adequately by reporting a single accuracy value. A system's output depends on image quality, finger placement, sensor technology, enrollment strategy, matching threshold, and the population represented in testing. As a result, performance figures from different studies may not be directly comparable unless datasets and protocols are aligned.

Performance evaluation is therefore a central part of biometric research. ISO/IEC 19795-1 specifies general principles for empirical biometric performance testing and reporting, including error rates, throughput, test protocols, and the need to reduce bias and clarify limits of applicability (International Organization for Standardization [ISO] & International Electrotechnical Commission [IEC], 2021). Fingerprint Verification Competition campaigns further demonstrated the value of common databases and controlled comparison protocols for tracking progress across algorithms (Cappelli 2006).

This paper focuses on how evaluation metrics and benchmark datasets jointly shape conclusions about fingerprint authentication. It emphasizes that meaningful evaluation requires a clearly stated task verification or identification and evidence across operating thresholds, data conditions, and failure modes.

II. RESEARCH OBJECTIVES

1. To explain the principal metrics used to evaluate fingerprint verification and identification.
2. To review commonly used benchmark datasets and describe their research value and limitations.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, September 2026
Impact Factor: 8.2

3. To identify protocol choices that affect reproducibility, fairness, and generalization.
4. To propose an integrated evaluation framework for fingerprint authentication research.

III. RESEARCH METHODOLOGY

This paper adopts a narrative and comparative literature-review approach. Foundational fingerprint evaluation studies, benchmark competition publications, standards documentation, and dataset catalogues were consulted. Sources were selected for their direct relevance to fingerprint matching, biometric testing principles, benchmark design, and reporting. The analysis organizes evidence by (a) task type, (b) metric family, (c) dataset characteristics, and (d) protocol and reporting requirements. No new biometric experiments or numerical performance claims are presented; tables summarize dataset descriptions and methodological guidance from the cited sources. Therefore, the paper's contribution is a synthesis and evaluation framework rather than an empirical ranking of algorithms.

IV. PERFORMANCE EVALUATION METRICS

1.Verification error rates

Fingerprint verification is a one-to-one comparison: a probe is compared with a claimed identity's enrolled template. A matcher produces a similarity score and accepts or rejects the claim at a threshold. FMR is the proportion of impostor comparison attempts incorrectly accepted, whereas FNMR is the proportion of genuine attempts incorrectly rejected. In some literature, the corresponding terms false accept rate and false reject rate are used; definitions and denominators should be stated because terminology can vary by test protocol. Lower FMR and FNMR are desirable, but they trade off against each other as the threshold changes (ISO/IEC, 2021).

2. Equal error rate and threshold selection

The equal error rate is the operating point at which FMR and FNMR are equal or approximately equal. EER is convenient for summarizing a matcher, but it is not a universal deployment metric: an application may require a much lower impostor acceptance rate, and the threshold producing that requirement can yield a substantially different genuine rejection rate. Studies should report EER alongside threshold-specific FMR/FNMR values and should avoid selecting a threshold using the test set.

3. ROC and DET curves

ROC curves plot genuine acceptance performance against impostor acceptance across thresholds; DET curves display error rates in a form that can make differences at low-error operating regions easier to inspect. Curves reveal whether one algorithm dominates another across thresholds or whether performance depends on the selected operating point. Reporting a curve, its underlying score distributions, and a selected threshold is more informative than a lone scalar.

4. Identification metrics

In one-to-many identification, a probe is searched against a gallery. Rank-1 identification rate measures the fraction of probes for which the correct identity is the highest-ranked candidate; cumulative match characteristic curves report recognition at multiple ranks. For open-set identification, where some probes are not enrolled, false-positive identification rate and detection/identification rate should also be reported. Verification metrics cannot be substituted for identification metrics without adapting the protocol.

5. Failure, quality, and efficiency measures

Failure-to-enroll and failure-to-acquire rates capture transactions that do not produce usable templates or samples. These failures should be included rather than silently discarded, because excluding poor-quality captures can inflate apparent accuracy. Image-quality scores, such as NIST Fingerprint Image Quality indicators, can support stratified analysis but do not replace matching performance. Throughput, latency, memory, and computational requirements are also relevant for embedded or high-volume deployments (ISO/IEC, 2021).

Metric	Meaning	Best interpretation / caution
FMR / FAR	Impostor attempts accepted	Report denominator and threshold; critical for security

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, September 2026 Impact Factor: 8.2

FNMR / FRR	Genuine attempts rejected	Reflects usability and capture/matching difficulty
EER	Threshold where FMR ≈ FNMR	Compact summary; may not match deployment target
ROC / DET	Error trade-off across thresholds	Supports operating-point comparison
Rank-1 / CMC	Correct identity ranking in gallery search	Identification, not verification
FTE / FTA	Enrollment/acquisition failures	Report separately and in end-to-end outcomes
Latency / throughput	Time or transactions per unit time	Include hardware, software, and test conditions

V. BENCHMARK FINGERPRINT DATASETS

Benchmark datasets provide shared samples, controlled conditions, and a basis for reproducibility. They differ in sensor type, resolution, sample count, capture sessions, image quality, and whether images are rolled, plain, latent, or synthetic. The FVC series is particularly influential in fingerprint verification research. The FVC2000, FVC2002, and FVC2004 campaigns provided structured databases and common evaluation procedures, enabling direct algorithm comparison under defined conditions (Cappelli 2006).

Dataset / resource	Typical content / design	Research value	Important limitation
FVC2000	Four subsets; 110 fingers × 8 impressions per subset as catalogued by IAPR TC4	Early shared verification benchmark; supports controlled comparisons	Older sensors and acquisition conditions; limited demographic and modern-device coverage
FVC2002	Four subsets; 100 fingers × 8 impressions per subset	Sensor diversity and standardized protocol	Small subject pool; possible overfitting to familiar benchmark
FVC2004	Four subsets; 100 fingers × 8 impressions per subset; real and synthetic data across subsets	Improved data variety and challenging comparisons	Synthetic and real subsets should be distinguished; not all conditions mirror deployment
NIST SD4	2,000 grayscale fingerprint image pairs, 512 × 512 pixels	Useful classic image collection for fingerprint processing and matching research	Dataset purpose and image characteristics differ from contemporary live-scan operational tests
Latent fingerprint datasets	Latent impressions, often partial, noisy, and forensic in origin	Tests difficult image conditions and forensic workflows	Not interchangeable with clean live-scan verification benchmarks

The FVC-onGoing platform extends benchmark practice through ongoing evaluation and sequestered datasets, allowing algorithm assessment beyond a one-time competition and supporting module-level testing (Biometric System Laboratory, University of Bologna, n.d.). NIST's database catalogue describes SD4 as a collection of 2,000 fingerprint image pairs, while newer NIST releases include latent and exemplar materials designed for forensic suitability research (Fiumara 2026).

1. Dataset selection principles

A benchmark should be selected according to the intended application. A mobile access-control system requires live-scan samples and sensor variation; a forensic workflow requires partial latent impressions and examiner-oriented conditions; an algorithm-development study may use controlled datasets but should validate on a held-out source. Dataset documentation should include subjects, fingers per subject, capture sessions, sensor make/model, resolution, image format, acquisition conditions, quality distribution, demographic coverage where ethically and legally appropriate, and licensing or access restrictions.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, September 2026
Impact Factor: 8.2

2. Avoiding data leakage

A major threat is leakage between training, validation, and test sets. Randomly splitting individual images can place impressions of the same finger in both training and testing, allowing the model to exploit identity-specific patterns. For credible generalization estimates, splits should be made at subject or finger identity level, with session- or sensor-disjoint testing when the research question concerns cross-session or cross-device transfer. Threshold tuning and model selection must use development data, not the final test set.

VI. INTEGRATED EVALUATION FRAMEWORK

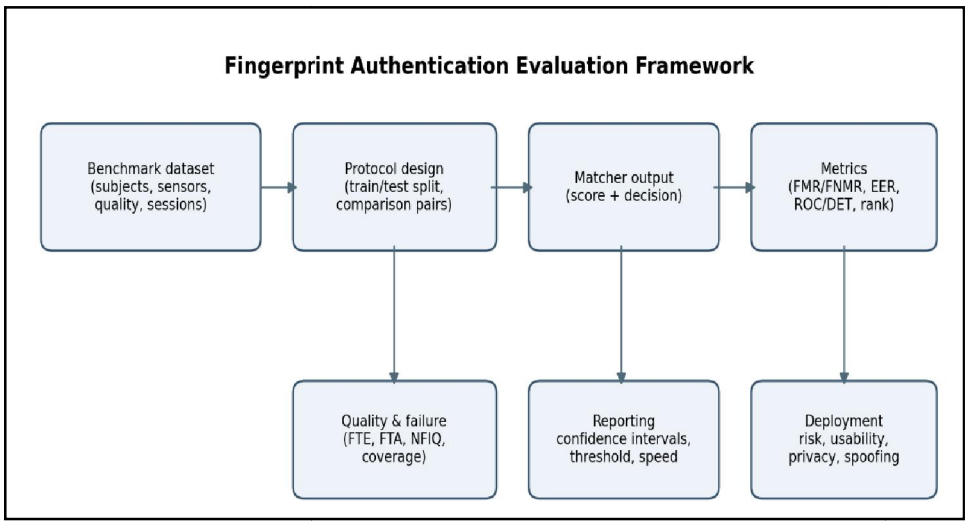


Figure 1. Conceptual framework linking benchmark composition, protocol design, matcher outputs, metrics, quality/failure reporting, and deployment interpretation.

Figure 1 illustrates that evaluation is a pipeline rather than a metric-only exercise. Dataset composition defines which conditions are represented; protocol design determines the validity of comparisons; matcher scores are converted into operating-point metrics; and quality, failures, uncertainty, and system-level factors inform deployment decisions.

VII. COMPARATIVE DISCUSSION

Metrics and datasets are complementary. Metrics define what is measured, while datasets define the conditions under which it is measured. EER provides a concise summary but can hide security/usability trade-offs; ROC/DET curves provide richer threshold information but still depend on representative samples. FVC databases offer controlled comparability, yet their relatively small and historically specific capture populations limit direct extrapolation to modern deployments. Latent datasets expose difficult forensic conditions but should not be used as a proxy for ordinary touch-sensor verification.

A benchmark can become a target in itself: repeated tuning on a public dataset may improve benchmark scores while reducing performance on unseen devices or populations. This risk motivates hidden test sets, cross-dataset validation, repeated capture sessions, and transparent reporting. ISO/IEC 19795-1 emphasizes protocol documentation and limits of applicability, which supports the principle that performance claims should be tied to a test population and transaction definition rather than presented as universal accuracy (ISO/IEC, 2021).

Evaluation dimension	Minimum recommended reporting
Data	Dataset name/version, subjects, fingers, impressions, sessions, sensor, image quality
Protocol	Task, genuine/impostor construction, gallery size, split strategy, threshold selection
Metrics	FMR/FNMR at operating points, EER, ROC/DET; CMC/rank metrics for identification

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, September 2026
Impact Factor: 8.2

Failures	FTE, FTA, excluded samples and reason; end-to-end success rate
Uncertainty	Confidence intervals or repeated-run variability; number of comparisons
Implementation	Algorithm version, preprocessing, hardware, software, runtime
Generalization	Cross-session, cross-sensor, cross-dataset or operational validation

VIII. RECOMMENDATIONS

1. Use ISO/IEC 19795-1 principles as a reporting baseline and clearly label the evaluation type (technology, scenario, or operational).
2. Report threshold-dependent FMR and FNMR alongside EER; select thresholds using development data and freeze them before testing.
3. Use verification metrics for verification and CMC/rank metrics for identification; define open-set measures when unenrolled probes are possible.
4. Include FTE, FTA, quality distributions, sample counts, and confidence intervals in end-to-end reporting.
5. Use multiple datasets and hold out subjects; add cross-session and cross-sensor tests to estimate generalization.
6. Publish complete protocol details, dataset versions, implementation settings, and code where licensing and privacy permit.
7. Treat presentation-attack resistance, privacy, usability, and fairness as complementary evaluation dimensions, not as substitutes for matcher accuracy.

IX. CONCLUSION

Performance evaluation metrics and benchmark datasets are essential to determining whether fingerprint-based biometric authentication is accurate, reliable, and suitable for a given application. FMR/FNMR and EER summarize verification errors, ROC/DET curves expose threshold trade-offs, and rank-based measures assess identification. Benchmark resources such as FVC and NIST collections provide shared foundations, but their value depends on careful protocol design and recognition of their limits. The most credible evaluation combines representative data, identity-disjoint splits, multiple operating points, quality and failure reporting, uncertainty estimates, and cross-condition validation. Accordingly, future fingerprint research should move beyond isolated accuracy claims toward transparent, reproducible, application-oriented evaluation that links laboratory results to real-world performance.

REFERENCES

1. Cappelli, R., Maio, D., Maltoni, D., Wayman, J. L., & Jain, A. K. (2006). Performance evaluation of fingerprint verification systems. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(1), 3–18.
2. Dorizzi, B., Chollet, G., & Petrovska-Delacrétaz, D. (Eds.). (2009). Guide to biometric reference systems and performance evaluation. Springer.
3. Fiumara, G., Ko, K., & Wertheim, K. (2026). NIST Special Database 303 (NIST Technical Note 2366). National Institute of Standards and Technology.
4. Garris, M. D., & Wilson, C. L. (2005). NIST biometric evaluations and developments (NISTIR 7204). National Institute of Standards and Technology.
5. Maio, D., Maltoni, D., Cappelli, R., Wayman, J. L., & Jain, A. K. (2002). FVC2000: Fingerprint verification competition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(3), 402–411.
6. Maio, D., Maltoni, D., Cappelli, R., Wayman, J. L., & Jain, A. K. (2002). FVC2002: Second fingerprint verification competition. Proceedings of the International Conference on Pattern Recognition, 811–814.
7. Maio, D., Maltoni, D., Cappelli, R., Wayman, J. L., & Jain, A. K. (2004). FVC2004: Third fingerprint verification competition. Proceedings of the International Conference on Biometric Authentication, 1–7.