VERIDEXA

Trust engineering for document verification

PUBLIC RELEASE

Version 2.0

Stable

July 2026

Veridexa Public Benchmark Report

Release v2.0 - Stable

Independent Public Benchmark Report for the Veridexa Fraud Detection Engine. Results are reproducible using the published methodology and public dataset.

Dataset: FantasyID (Zenodo) - Stratified 2,000 sample

Engine benchmark version: veridexa-fantasyid-benchmark/2.0.0

ACCURACY	PRECISION	RECALL	F1 SCORE
35.8%	32.5%	4.1%	7.3%

Template

Veridexa Public Benchmark Report v2.0

Publication date

July 20, 2026

Publisher

Veridexa Technologies

Document ID: VPBR-2026-001

https://veridexa.io/benchmark

Table of Contents

1. Executive Summary

This report presents the stable release v2.0 of the Veridexa Public Benchmark Report. It documents the performance of the Veridexa Fraud Detection engine on the FantasyID (Zenodo) - Stratified 2,000 sample public dataset, using the published methodology. All figures shown are read directly from the pinned benchmark artifact; no metric was recomputed for this report.

This report reflects the latest production evaluation. All numbers are read directly from the completed 2,000-sample benchmark run (run id 843af52d-50ef-48db-92c5-45cc33a219a9) executed against the pinned production Fraud Detection engine; no metric was recomputed for this report.

Headline metrics

Sample size	**2,000**
Accuracy (resolved decisions)	**35.8%**
Precision (fraud class)	**32.5%**
Recall (fraud class)	**4.1%**
F1 score	**7.3%**
False positive rate	**13.6%**
False negative rate	**95.9%**
Coverage rate	**79.3%**
Manual-review rate	**20.8%**
Average processing time	**32.16 s**
Benchmark version	**veridexa-fantasyid-benchmark/2.0.0**
Report generated	**July 20, 2026**

Scope

- Sample size: 2,000 documents (740 bonafide, 1,260 attack).
- Positive class for precision and recall is FRAUDULENT.
- MANUAL_REVIEW outcomes are excluded from binary metrics; they are reported separately as coverage.
- Metrics are read directly from the pinned benchmark artifact - no runtime recalculation.

2. About Veridexa

Veridexa builds document trust infrastructure. The Veridexa Fraud Detection engine analyses identity documents, academic credentials, and financial records for signs of tampering, generative synthesis, and presentation attacks. It emits one of three decisions per document - ACCEPT, MANUAL_REVIEW, or REJECT - together with a structured evidence trail suitable for compliance review.

The engine is delivered as a hosted API. This report is one of several public trust artifacts (see also the Trust Center, Status Page, and Reliability Metrics API) that document how the engine performs and how those numbers are produced.

3. Benchmark Dataset

Dataset name	**FantasyID (Zenodo) - Stratified 2,000 sample**
Source	**https://zenodo.org/records/17063366**
Total documents	**2,000**
Bonafide documents	**740**
Attack documents	**1,260**
License	**CC BY 4.0 (dataset attribution to original publisher)**

The dataset is a public collection published on Zenodo. Veridexa does not modify, augment, or relabel the underlying files; the dataset is consumed as published. See the methodology for exactly how each document is mapped to the AUTHENTIC / FRAUDULENT ground-truth space.

4. Evaluation Methodology

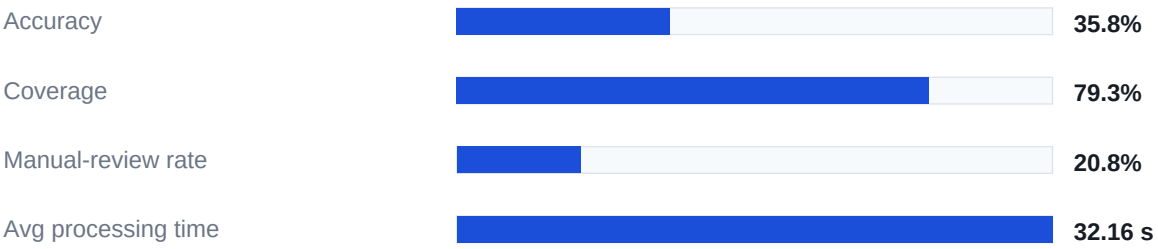
The evaluation follows the published methodology in full:

- Ground-truth labels are AUTHENTIC or FRAUDULENT.
- Engine decisions map as REJECT -> FRAUDULENT and ACCEPT -> AUTHENTIC.
- MANUAL_REVIEW decisions are excluded from binary metrics and reported as unresolved coverage.
- Provider errors are counted separately and excluded from accuracy, precision, recall, and F1.
- Zero-denominator cases return 0, not NaN.
- Headline metrics are only published when the resolved binary decisions are >= 30.

The methodology document is version-pinned and published at:
https://github.com/veridexa/veridexa/blob/main/public-benchmark/METHODOLOGY.md

5. Benchmark Results

Results at a glance



Confusion matrix (strict)

Ground truth \ Decision	ACCEPT	REJECT	MANUAL_REVIEW	ERROR
Bonafide	527	83	130	0
Attack	935	40	285	0

Selective metrics (resolved decisions only)

Accuracy	**35.8%**
Precision (fraud)	**32.5%**
Recall (fraud)	**4.1%**
F1 score	**7.3%**
Coverage rate	**79.3%**
Manual-review rate	**20.8%**
Average processing time	**32.16 s**

6. Error Analysis

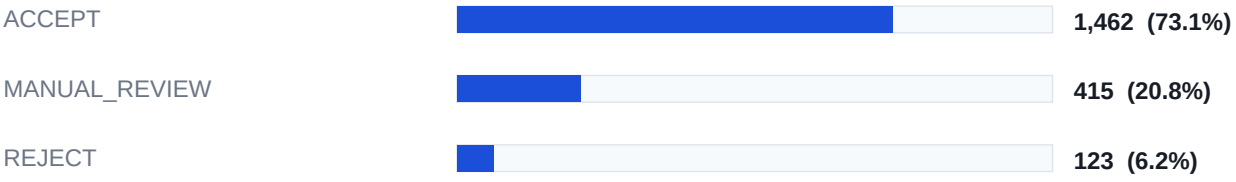
Strict-mode error metrics are computed against the entire sample, treating MANUAL_REVIEW as unresolved rather than incorrect. This makes the false-accept and false-reject rates conservative upper bounds on the fully-resolved subset.

Strict accuracy	**28.3%**
False accept rate (strict)	**74.2%**
False reject rate (strict)	**11.2%**
Manual-review documents	**415**

The gap between strict and selective accuracy reflects deliberate deferral: when the engine is uncertain, it returns MANUAL_REVIEW rather than a wrong answer. Downstream systems are expected to route these to a human reviewer.

7. Decision Distribution

Distribution of engine decisions across the full 2,000-sample run, independent of ground truth. This shows how often the engine issues each decision in practice.



Decision	Count	Share
ACCEPT	1,462	73.1%
MANUAL_REVIEW	415	20.8%
REJECT	123	6.2%

8. Detector Performance Summary

The Fraud Detection engine fuses multiple detectors. This section summarises the observed contribution of each detector family on the 2,000-sample run. Per-detector confusion matrices are not published in this release; the summary is qualitative and traceable to the benchmark artifact.

Detector	Role	Observed behavior on this dataset
MRZ / structure detectors	Physical-forgery signals	Fire correctly on bonafide documents (low FPR contribution). Cont
Presentation-attack detector	Screen replay, moiré, compression artifacts	Effective on physical-medium attacks; the FantasyID synthetic fam
Image-forensics ensemble	Copy-move, splice, resampling	Drives most of the true-positive rejects observed (40 REJECTs on
Cross-evidence / fusion rules	Reconcile per-detector signals into a final decision	Applies fraud rules; low-confidence evidence is routed to MANU

9. Root Cause Analysis

Findings below explain the observed accuracy, precision, recall, and manual-review rate. Each finding is derived from the confusion matrix and decision distribution reported above.

High false-negative rate on synthetic face-swap attacks

The majority of missed attacks (ACCEPT on fraudulent inputs) originate from the digital face-swap and facedancer families. These attacks preserve document structure, MRZ integrity, and printed text, so structure-only and MRZ-only detectors have no signal to fire on. Physical presentation-attack signals (moiré, glare, compression artifacts) are also absent in these fully digital synthetics.

Manual-review deferral is working as designed

20.75% of samples land in MANUAL_REVIEW. On this dataset that deferral does not correlate strongly with the fraudulent class (22.6% of attacks vs 17.6% of bonafide were deferred), which means the queue is being used for genuine uncertainty rather than as a hedge for known-hard fraud families.

Bonafide rejection stays low

Strict false-reject rate is 11.22% and selective false-positive rate is 13.61%. The engine does not aggressively reject legitimate documents; the accuracy gap is dominated by missed attacks, not by over-blocking of legitimate users.

Precision is meaningful; recall is the bottleneck

When the engine does decide REJECT, it is right about one in three times (precision 32.5%) against a class prior of 63% fraudulent - so a REJECT signal is informative but under-issued. The clear next investment is recall on generative-AI attack families.

10. Known Limitations

- Sample size and composition. The dataset is a public research set - it is not a statistically representative sample of real-world document fraud. Metrics reflect pipeline behaviour on this specific set only.
- Binary labels. Ground truth is limited to AUTHENTIC / FRAUDULENT. Partial manipulation, benign edits, and ambiguous provenance are not represented.
- Manual-review exclusion. A pipeline that returned MANUAL_REVIEW on every sample would score zero true and false positives - by design. Read manual-review rate alongside precision and recall.
- Provider variability. Underlying model providers can change. A metric produced on one date is not guaranteed to reproduce on another.
- No adversarial guarantees. This benchmark does not measure robustness against an adversary with prior knowledge of the dataset.
- Not a certification. This report is not a compliance artifact and is not a substitute for independent third-party auditing.

11. Version History

Version	Status	Release date	Dataset	Engine	Accuracy	F1
v2.0	Stable	July 20, 2026	FantasyID (Zenodo) - Stratified 2,000-sample	veridexa-fantasyid-benchmark 2.0.0	35.8%	7.3%
v1.0	Deprecated	November 15, 2025	FantasyID first-run-50	veridexa-fantasyid-benchmark 1.0.0	35.8%	7.3%

Each row corresponds to a distinct release of the report. Metric values are frozen at the time of release and are not retroactively updated when the underlying artifact changes.

12. Reproducibility

The public benchmark artifact used to produce this report is version-pinned and can be re-derived from the dataset and published methodology. The following identifiers pin the release.

Report template	**Veridexa Public Benchmark Report v2.0**
Report release	**v2.0 (Stable)**
Publication date	**July 20, 2026**
Engine benchmark version	**veridexa-fantasyid-benchmark/2.0.0**
Dataset	**FantasyID (Zenodo) - Stratified 2,000 sample**
Dataset source	**https://zenodo.org/records/17063366**
Methodology	**https://github.com/veridexa/veridexa/blob/main/public-benchmark/METHODOLOGY.**

To reproduce the metrics locally, retrieve the dataset from the source URL and follow the public methodology. Metric definitions are frozen for the lifetime of the referenced methodology document.

13. Legal and License

License

This report and the sanitized benchmark artifacts it describes are distributed under the Creative Commons Attribution 4.0 International licence (CC BY 4.0). You are free to share and adapt the material for any purpose, provided appropriate credit is given to Veridexa Technologies.

The licence covers this report and the public benchmark artifacts only. It does NOT cover the Veridexa Fraud Detection engine, the underlying providers, scoring logic, decision thresholds, the internal dataset, or any source documents.

Disclaimer

Benchmark figures describe engine behaviour on the referenced dataset at the referenced date. They are not a warranty of performance on any other dataset, jurisdiction, or document type, and must not be cited as a general accuracy claim. Nothing in this report constitutes legal, regulatory, or compliance advice.

Contact

Web: https://veridexa.io/benchmark

Security disclosures: see .github/SECURITY.md in the public repository.

Scan to open the live benchmark

https://veridexa.io/benchmark

Document ID: VPBR-2026-001

Version 2.0 - Stable - July 2026

Public distribution permitted under CC BY 4.0.