

Paper Type: Original Article

An Integrated AI-Based Forensic Analysis Platform for Multimodal Biometric, Genetic, and Digital Evidence

Anas Tamer ¹ , Abdelrahman Sherif ¹ , Nada Wafai ¹ , Tasneem Sakr ¹ , Taghreed Mohamed ¹ ,
Ahmed AbdelQuder ¹ , and Ahmed Hossam ^{1,*} 

¹ Faculty of Computers and Informatics, Zagazig University, Zagazig 44519, Sharqiyah, Egypt;
Emails: anastamer872004@gmail.com, Abdelrahmansherif1106@gmail.com, nadaelkafrawy@icloud.com,
tasneemsakr87@gmail.com, taghreedbasuny6@gmail.com, anaahmedmohamed48@gmail.com, adarwish@fci.zu.edu.eg.

Received: 07 May 2026**Revised:** 06 Aug 2026**Accepted:** 01 Oct 2026**Published:** 07 Oct 2026

Abstract

Forensic investigations increasingly combine visual, biometric, genetic, and digital evidence, yet the analytical tools used for these modalities are commonly separated from case management and evidence handling. This study presents a modular forensic analysis platform that integrates four AI-assisted functions - deepfake screening, face recognition, SNP-based phenotype analysis, and face restoration with downstream demographic estimation - within a case-centric web and mobile workflow. The scientific contribution is architectural and workflow-level: established analytical models are orchestrated through a common evidence lifecycle so that an authorized practitioner can register a case, attach evidence, invoke the appropriate analytical service, and retain the resulting output with the case record. The reported internal evaluation is intentionally module-specific. The deepfake classifier achieved 99.52% accuracy on a balanced 20,000-image test set (19,904 correct predictions); the five-identity face-recognition integration test achieved 94.55% accuracy (52/55); Gradient Boosting produced the highest reported accuracy in the encoded SNP-classification benchmark at 98.34%; and the downstream demographic stage achieved 87% accuracy for the dataset's gender label and 56% for age-group classification. Face restoration is demonstrated qualitatively because no paired-reference restoration metric was recorded in the available experiment. These results are not treated as directly comparable across tasks or as evidence of universal forensic validity. Instead, they show that heterogeneous analytical modules can be combined into one controlled workflow while preserving task-specific interpretation. The platform is positioned as a decision-support environment in which algorithmic outputs require professional review rather than as an autonomous evidentiary decision maker.

Keywords: Digital Forensics; Forensic Artificial Intelligence; Deepfake Detection; Face Recognition; Forensic DNA Phenotyping; Image Restoration; Case Management; Multimodal Evidence.

1 | Introduction

Modern forensic practice is increasingly multimodal. A single investigation may contain images or video requiring authenticity assessment, facial material requiring identity comparison, biological traces suitable for genotype-based inference, and degraded imagery that must be restored before a human examiner can inspect it. Research in each of these areas has advanced rapidly: manipulated-face detection has been studied on large-



Corresponding Author: adarwish@fci.zu.edu.eg



<https://doi.org/10.61356/j.scin.2026.3679>



Licensee **SciNexuses**. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

scale benchmarks such as FaceForensics++ and the DeepFake Detection Challenge [1, 2]; residual convolutional networks provide strong visual representations [3]; RetinaFace and ArcFace are widely used for face localization and discriminative embedding generation [4, 5]; and forensic DNA phenotyping systems such as IrisPlex and HIrisPlex-S demonstrate how informative SNPs can support probabilistic inference of externally visible characteristics [6, 7].

The operational problem is different from the model-development problem. A laboratory or investigative team must move evidence through a controlled sequence of registration, storage, analysis, review, and reporting. When each analytical capability is delivered as a separate application, users must repeatedly transfer files, maintain parallel case identifiers, and manually reconnect model outputs to their source evidence. This increases workflow complexity and makes it harder to preserve a coherent analytical record. The present work therefore focuses on system-level integration rather than claiming a new deepfake detector, face-recognition loss, restoration network, or DNA-phenotyping theory.

1.1 | Research Objective and Scientific Contribution

The objective is to design and evaluate a case-centric forensic platform in which heterogeneous analytical services are exposed through one authenticated workflow. The contribution is fourfold. First, the platform defines a common evidence-to-analysis pathway for visual, biometric, and genetic inputs instead of presenting the models as unrelated demonstrations. Second, the service architecture separates authentication, case management, file handling, analytical inference, and persistence so that the system can evolve without coupling the user interface to a single model. Third, the study reports module-specific experimental results without collapsing heterogeneous tasks into one misleading global accuracy value. Fourth, the system design explicitly places AI output inside a human-reviewed forensic workflow; model predictions are treated as investigative support and remain distinguishable from examiner conclusions.

This distinction is important for scientific positioning. The originality of the work is the orchestration of multiple established analytical methods around a shared case and evidence model, together with a practical web/mobile implementation and an evaluation that exposes both strong and weak modules. The paper therefore avoids unsupported state-of-the-art claims and evaluates each component only against the evidence available from its own experiment.

2 | Related Work and Functional Positioning

2.1 | Task-Specific Forensic AI

Deepfake detection remains sensitive to dataset composition, compression, generation method, and domain shift. FaceForensics++ showed the value of standardized manipulation benchmarks, while the DFDC dataset expanded the scale and diversity of face-swap data [1, 2]. These studies motivate the use of a separate held-out test set and caution against interpreting high within-dataset accuracy as universal robustness. For the present prototype, ResNet-50 is used as the visual backbone because residual learning supports effective optimization of deep convolutional networks [3].

For facial identity analysis, the system combines face localization/alignment with embedding-based recognition. RetinaFace provides dense face localization and facial landmarks [4], while ArcFace learns angularly separated embeddings that are well suited to similarity-based verification [5]. In forensic use, however, image quality and demographic effects remain material concerns. NIST evaluations have documented that face-recognition error rates can vary with image quality and demographic group, so a candidate match should be reviewed as an investigative lead rather than accepted automatically [14].

Forensic DNA phenotyping is conceptually different from biometric identification. IrisPlex predicts eye color from a small set of informative SNPs, and HIrisPlex-S extends trait prediction to eye, hair, and skin color [6, 7]. SNP metadata and flanking sequence information can be obtained from resources such as NCBI dbSNP

[8]. These systems support intelligence generation when standard identity matching is unavailable, but the outputs are probabilistic trait estimates rather than unique identification.

For degraded facial imagery, the platform uses pretrained restoration and super-resolution components. GFP-GAN uses generative facial priors for blind face restoration [9], while Real-ESRGAN targets practical blind super-resolution under complex degradations [10]. The FFHQ dataset introduced with StyleGAN provides a large high-quality face corpus that has been widely used in face-generation and restoration research [11]. In this study, restoration and downstream demographic estimation are treated as separate outputs so that enhancement quality is not inferred from classification accuracy alone.

2.2 | Integrated Forensic Platforms

Existing forensic products demonstrate strong capabilities within their own scope. Autopsy provides case management and extensible digital-forensics analysis [15], while Magnet AXIOM integrates acquisition, artifact analysis, and case-centric investigation of digital sources [16]. Clearview AI publicly describes a law-enforcement-oriented facial search service [17], and Parabon Snapshot provides DNA phenotyping and related genetic intelligence services [18]. These systems are not directly comparable products, and the public documentation does not establish that one system is globally superior to another. Table 1 is therefore a functional positioning table, not a performance ranking.

Table 1. Functional positioning based on capabilities documented in the cited public sources.

System	Publicly documented primary scope	Case/evidence workflow	Face identity	Deepfake screening	DNA phenotyping	Face restoration	Integration scope
Autopsy [15]	Digital-device/file forensics	Yes	N/D	N/D	N/D	N/D	Digital evidence centered
Magnet AXIOM [16]	Digital evidence recovery and analysis	Yes	N/D	N/D	N/D	N/D	Digital evidence centered
Clearview AI [17]	Facial search for investigations	Limited/public scope not case-workflow focused	Yes	N/D	N/D	N/D	Single primary biometric function
Parabon Snapshot [18]	Forensic DNA analysis and phenotyping	Service-specific	N/D	N/D	Yes	N/D	Genetic analysis centered
Proposed platform	Multimodal forensic decision support	Yes	Yes	Yes	Yes	Yes	Case-centric multimodal workflow

N/D = not documented in the cited public product scope. N/D is not proof that a capability is absent. The table compares functional scope only; it is not a benchmark of accuracy, legal admissibility, or product quality.

Table 1 makes the integration gap explicit. General digital-forensics suites provide rich case and device-analysis capabilities, whereas task-specific services concentrate on facial search or genetic analysis. The proposed platform's contribution is the co-location of these heterogeneous functions under one case record. This breadth should not be interpreted as replacement of specialized laboratory systems; its purpose is workflow unification and coordinated decision support.

3 | System Architecture and Workflow Design

3.1 | Service-Oriented Architecture

The platform is organized as a layered, service-oriented architecture. Figure 1 shows the implementation structure: web and mobile clients communicate through an API gateway; authentication and role-based access control are separated from core services; case, user, AI, and file operations are represented as distinct services; and persistence is divided across user, case, AI-result, and file-metadata stores. Caching, object storage, and logging/monitoring are represented as supporting infrastructure. This separation reduces direct coupling between the user interface and individual analytical models and allows a model service to be replaced without redesigning the entire application.

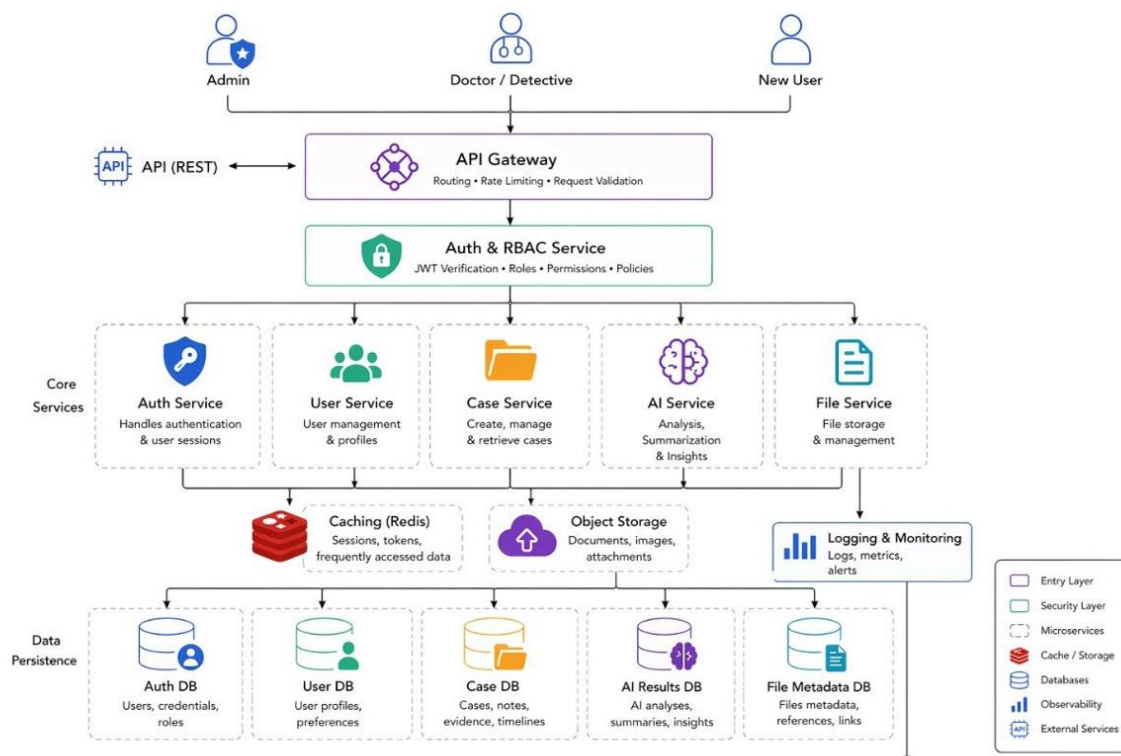


Figure 1. Software architecture of the proposed platform, showing the API gateway, authorization layer, core services, supporting infrastructure, and persistence layer.

3.2 | Evidence-Centric Analysis Lifecycle

The unit of work is the case rather than the model. An authorized practitioner first selects or creates a case, attaches an evidence item, chooses an analytical service, and runs inference. The output is then associated with the case so that the analysis can be revisited with its source evidence. This design avoids the common prototype pattern in which a model produces a result that is disconnected from the case context. The same interaction pattern is used for deepfake screening, face recognition, DNA analysis, and image restoration.

The platform supports three broad user paths: administrative management, professional forensic use, and public awareness/education. The analytical services are intended for authorized professional use; public-facing content is separated from case analysis. The community and chatbot components are supportive functions rather than evidentiary engines. In particular, chatbot responses should not replace examiner judgment, laboratory interpretation, or formal reporting.

3.3 | Technology Stack and Access Control

The web client is implemented with Angular and RxJS, and the mobile client uses React Native with Expo. Laravel provides the backend API layer, with Laravel Sanctum used for authenticated token-based access. MySQL stores structured application data, while file/object storage is used for evidence attachments and generated artifacts. The AI layer exposes model-specific inference endpoints. Development and integration used standard engineering tools for version control, API testing, dataset experimentation, and interface design.

Role-based access control is central to the design because forensic evidence is not ordinary application content. Administrative functions, professional case analysis, and general public access are separated at the authorization layer. The architecture also includes logging and monitoring components. These controls provide application-level traceability, but they should not be confused with a complete legal chain-of-custody mechanism. Courtroom admissibility depends on jurisdiction, laboratory procedure, evidence hashing and preservation policies, examiner documentation, and independent validation beyond the software architecture evaluated here.

4 | Materials and Methods

4.1 | Deepfake Detection

The deepfake module uses the 140K Real and Fake Faces dataset, which contains 70,000 real faces derived from Flickr imagery and 70,000 GAN-generated faces [12]. The dataset is distributed with balanced train, validation, and test partitions. Supplementary video-frame material was also used during development to expose the pipeline to frame-based inputs; because the available quantitative confusion matrix corresponds to a balanced 20,000-image test set, the reported performance in Section 5 is restricted to that clearly defined test set.

Images are converted to RGB, resized to the network input resolution, normalized, and augmented during training using transformations such as horizontal flipping and limited rotation. The classifier uses a pretrained ResNet-50 backbone with a binary output layer and transfer learning. The recorded training configuration uses a batch size of 64 and an initial learning rate of 1×10^{-4} . The model learns discriminative image representations directly; no claim is made that it explicitly measures a fixed handcrafted set of frequency or compression artifacts.

4.2 | Face Detection and Recognition

The face-recognition pipeline begins with RGB conversion, RetinaFace detection, landmark extraction, face-quality filtering, alignment, and cropping. Each accepted face is represented by a 512-dimensional embedding using an ArcFace-based recognition model [4], [5]. Identity comparison is performed with cosine similarity in the embedding space. The implementation was developed from an earlier small multiclass classifier toward an embedding-based design so that adding an identity does not require retraining a final classification layer.

The available quantitative experiment contains five identities and 55 test images. It is therefore treated as an integration-level recognition experiment, not as a large-scale biometric benchmark. This distinction prevents the small closed-set experiment from being presented as evidence of performance on large galleries or uncontrolled operational populations.

4.3 | SNP Processing and DNA Phenotype Analysis

The DNA module operates on a curated panel of 41 SNP loci associated with externally visible characteristics. Reference positions, alleles, and flanking sequence information are derived from the human reference context and NCBI dbSNP [8]. Input sequences are normalized, invalid symbols are rejected, strand orientation is handled, candidate loci are located through the stored flanking context, and observed alleles are converted

into encoded features. The trait interpretation is informed by the forensic DNA phenotyping literature, including IrisPlex and HIrisPlex-S [6], [7].

Two distinct experimental concepts are kept separate in this manuscript. The trait-level interface expresses eye, hair, and skin-color estimates, consistent with the phenotyping literature. The model-selection experiment available from the implementation, however, compares seven machine-learning classifiers on an encoded multiclass SNP dataset. Because the numeric classes in that benchmark are not equivalent to the published IrisPlex/HIrisPlex-S output categories, the classifier benchmark is reported as an encoded SNP-classification experiment and is not presented as a new validation of IrisPlex or HIrisPlex-S. This separation resolves an otherwise misleading conflation between a literature-based phenotyping model and an internal machine-learning benchmark.

4.4 | Face Restoration and Downstream Demographic Estimation

The restoration pipeline integrates GFP-GAN for face restoration and Real-ESRGAN for super-resolution [9], [10]. These models are used in pretrained form; the contribution of the present work is their integration into the case workflow rather than retraining the restoration networks. The face-restoration stage detects and aligns facial regions, restores them, and blends the enhanced regions back into the image. FFHQ is the principal high-quality face dataset associated with the generative facial prior used by GFP-GAN [11].

After restoration, a downstream demographic classifier produces the age-group and gender labels represented in the prototype interface. UTKFace provides more than 20,000 in-the-wild face images with age, gender, and ethnicity annotations and serves as a relevant annotated source for this stage [13]. The key methodological point is that demographic-classification accuracy is not a restoration-quality metric. A sharper image may still alter identity-relevant detail, and a classification result cannot substitute for reference-based measures such as PSNR, SSIM, perceptual similarity, or examiner assessment. Accordingly, Section 5 reports restoration qualitatively and demographic classification quantitatively.

4.5 | Module Inputs, Outputs, and Evaluation Basis

Table 2 summarizes the relationship between each analytical module, its primary input and output, and the evaluation evidence available in the present study. The table is included to prevent metrics from different tasks being interpreted as directly interchangeable.

Table 2. Analytical modules and the evidence used for evaluation in this study.

Module	Input	Core analytical path	Output	Evaluation basis
Deepfake screening	Face image or extracted frame	ResNet-50 binary classifier	Real/fake label and score	Balanced 20,000-image held-out test confusion matrix
Face identity	Detected/aligned face	RetinaFace + ArcFace-style embedding	Identity candidate / similarity	Five-identity, 55-image integration test
DNA analysis	DNA sequence / encoded SNPs	SNP parsing + encoded classifier benchmark; trait interpretation informed by FDP models	Encoded class and trait-level phenotype estimates	Seven-classifier internal benchmark plus trait-oriented interface
Restoration + demographics	Degraded facial image	GFP-GAN + Real-ESRGAN + downstream classifier	Restored image, age-group and gender labels	Qualitative restoration; 4,736-sample demographic classification report

As Table 2 indicates, the modules answer different forensic questions. Deepfake detection is a binary authenticity-screening task; face recognition is an identity-comparison task; the DNA component operates

on genetic features; and restoration changes image quality before a separate demographic classifier is applied. A single aggregate score across these tasks would therefore be scientifically meaningless.

4.6 | Evaluation Metrics and Interpretation Rules

Classification performance is reported using accuracy, precision, recall, and F1-score as available in the recorded classification reports. Confusion matrices are retained because they reveal the distribution of errors and class imbalance more clearly than a single scalar. Where the confusion matrix provides exact counts, the manuscript reports those counts rather than relying only on rounded percentages. No numerical comparison with external papers is used to claim superiority because the datasets, splits, class definitions, and operating points differ.

5 | Results

5.1 | Overall Module-Level Results

The principal quantitative results are summarized in Table 3. Each row refers to a different task and test population, so the values must be interpreted within the corresponding module rather than compared as if they came from one benchmark.

Table 3. Reported quantitative results by module.

Task	Class structure	Test size	Accuracy	Precision	Recall	F1-score	Interpretation
Deepfake detection	Binary real/fake	20,000 images	99.52%	99.52%	99.52%	99.52%	19,904/20,000 correct
Face recognition integration test	5 identities	55 images	94.55%	0.95 macro	0.95 macro	0.95 macro / 0.94 weighted	52/55 correct
Encoded SNP classification (Gradient Boosting)	10 encoded classes	Internal benchmark	98.34%	96.00%	97.34%	96.61%	Best of seven tested classifiers
Gender-label classification after restoration pipeline	2 labels	4,736 images	87%	87% macro	87% macro	87% macro	Dataset label terminology retained
Age-group classification after restoration pipeline	7 age groups	4,736 images	56%	54% macro	48% macro	50% macro	Substantial neighboring-class confusion

The SNP precision/recall/F1 values are those reported in the classifier-comparison output. The modules use different datasets and cannot be ranked by these percentages alone.

Table 3 shows why a reviewer should not treat the largest percentage as the strongest forensic evidence. The deepfake test is large and balanced but remains an in-dataset evaluation; the face test has only 55 images; the SNP benchmark uses encoded classes whose semantics differ from the trait-level interface; and the age classifier is markedly weaker than the gender-label classifier. Reporting these distinctions is more informative than presenting a single optimistic system accuracy.

5.2 | Deepfake Detection

Figure 2 presents the deepfake confusion matrix and classification report. The test set contains 10,000 real and 10,000 fake images. The model correctly classified 9,952 samples in each class and misclassified 48 in each direction, giving 19,904 correct predictions out of 20,000 and an accuracy of 99.52%. The symmetry of the two off-diagonal counts also explains why precision, recall, and F1-score are equal at 99.52% for both classes.

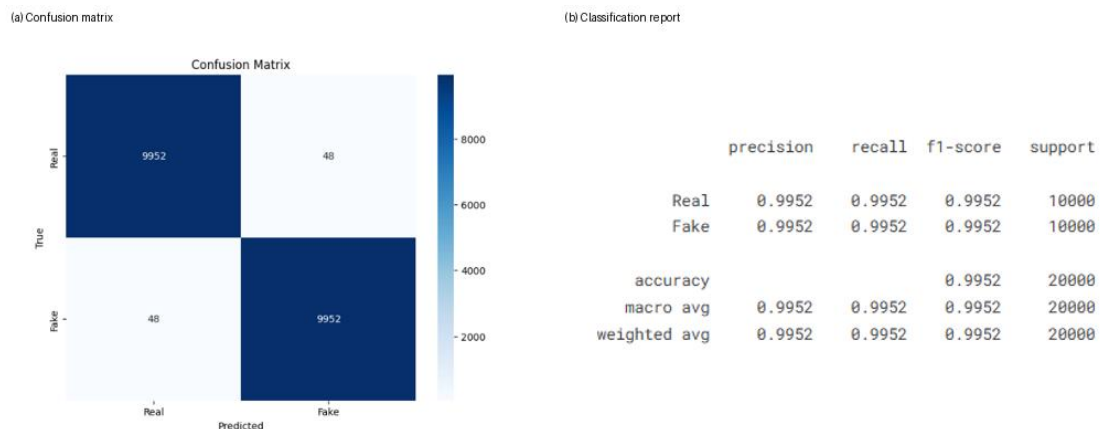


Figure 2. Deepfake test results: (a) confusion matrix and (b) classification report for the balanced 20,000-image test set.

This is a strong internal result, but its scope is deliberately narrow. Deepfake detectors can degrade when the manipulation method, compression pipeline, camera source, or post-processing differs from the training distribution [1], [2]. The result therefore supports the correctness of the implemented pipeline on the recorded test distribution; it does not establish universal deepfake detection performance.

5.3 | Face Recognition Integration Test

The five-identity evaluation in Figure 3 contains 55 images. The diagonal of the confusion matrix contains 52 correct classifications, corresponding to 94.55% exact accuracy. Three errors occur: two images of Chris Evans are assigned to other enrolled identities, and one image of Chris Hemsworth is assigned to Mark Ruffalo. The accompanying report gives macro precision, recall, and F1-score of approximately 0.95, with weighted F1-score of 0.94.

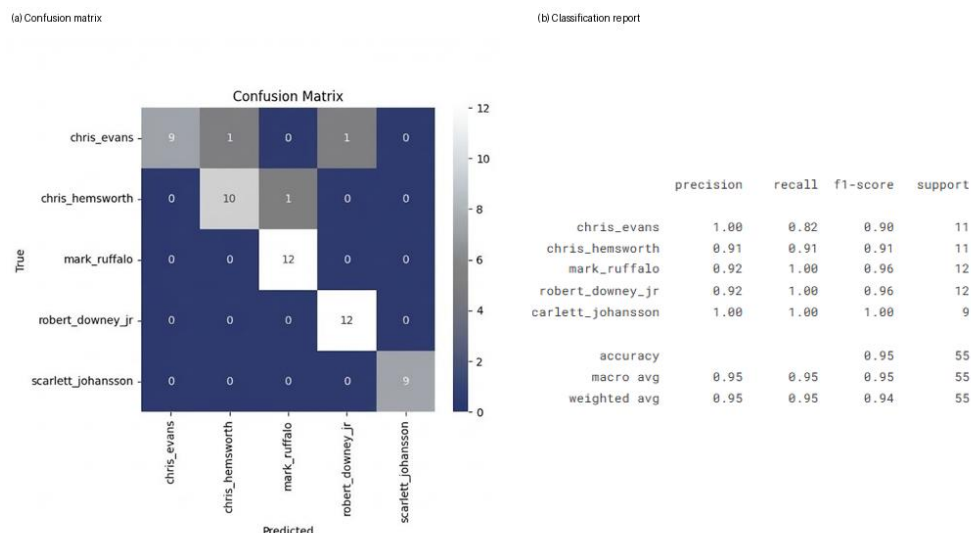


Figure 3. Five-identity face-recognition integration test: (a) confusion matrix and (b) per-identity classification report.

The small sample size is important to interpretation. The experiment demonstrates that the face preprocessing, identity representation, and application workflow operate coherently on the enrolled set, but it is not an estimate of error rates for a national-scale gallery or an uncontrolled forensic population. Operational deployment should additionally evaluate image quality, threshold-dependent false-match and false-non-match rates, demographic effects, and gallery size, consistent with the concerns emphasized in NIST face-recognition evaluations [14].

5.4 | Encoded SNP Classification Benchmark

Table 4 reports the seven classifiers compared in the encoded SNP experiment. Gradient Boosting achieved the highest reported accuracy and F1-score among the tested methods, followed closely by Random Forest and Decision Tree.

Table 4. Classifier comparison in the encoded SNP benchmark.

Classifier	Accuracy	Precision	Recall	F1-score
Gradient Boosting	98.3403%	96.0014%	97.3362%	96.6080%
Random Forest	97.8535%	96.5040%	94.6174%	95.3021%
Decision Tree	97.7871%	95.7221%	94.9390%	95.2735%
Logistic Regression	92.3434%	72.8225%	67.3044%	69.1297%
K-Nearest Neighbors	84.0009%	71.7550%	67.5280%	68.8731%
Naive Bayes	71.7858%	48.5762%	68.3324%	51.9784%
Support Vector Machine	65.5233%	18.4063%	20.8640%	19.5433%

Table 4 indicates a clear separation between the tree-based ensemble methods and the weaker baseline configurations in this experiment. The Gradient Boosting model reaches 98.34% accuracy with an F1-score of 96.61%, whereas the reported SVM configuration reaches 65.52% accuracy and a 19.54% F1-score. The disparity between accuracy and the lower precision/recall values for some classifiers also indicates that class frequency and multiclass error structure matter; accuracy alone is not an adequate summary. Figure 4 shows the Gradient Boosting confusion matrix. Most counts lie on the main diagonal, but the off-diagonal errors are not uniformly distributed. The figure therefore supports the classifier comparison while also showing that some encoded classes are easier to distinguish than others. Because the benchmark labels are numeric encoded classes, the manuscript does not relabel those classes as eye, hair, or skin categories without a verified mapping.

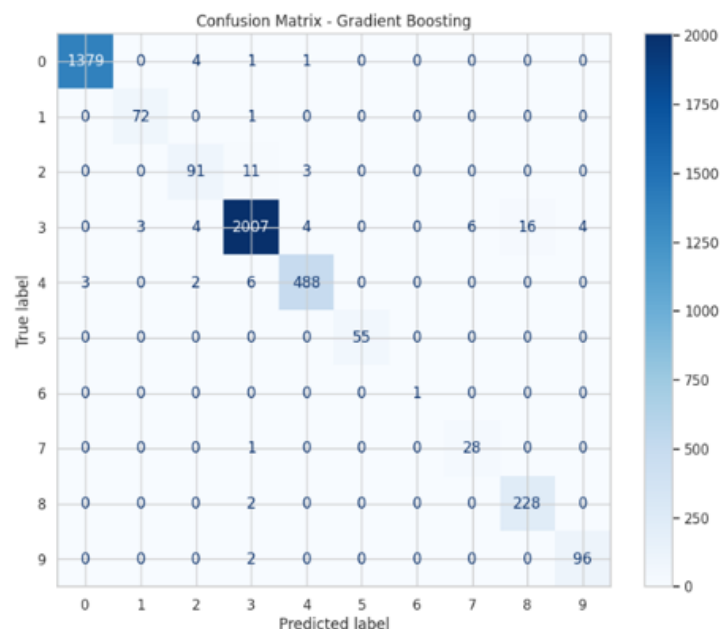


Figure 4. Confusion matrix for the Gradient Boosting model in the encoded SNP-classification benchmark.

5.5 | Restoration and Downstream Demographic Classification

The restoration pipeline is visually demonstrated in the application workflow, where a degraded facial image is passed through the enhancement service and returned as a clearer face image. The available experiment does not include paired ground-truth images or a recorded PSNR, SSIM, LPIPS, or identity-preservation score. For that reason, the manuscript does not convert the visual improvement into a numerical restoration claim. The downstream demographic evaluation is quantitative. Figure 5 shows the age-group confusion matrix together with the gender-label and age-group classification reports. The gender-label classifier achieves 87% accuracy on 4,736 images, with macro precision, recall, and F1-score all reported as 0.87. Age-group classification is more difficult: the overall accuracy is 56%, macro precision is 0.54, macro recall is 0.48, and macro F1-score is 0.50. The age matrix shows substantial confusion between neighboring groups, particularly around the large central classes.

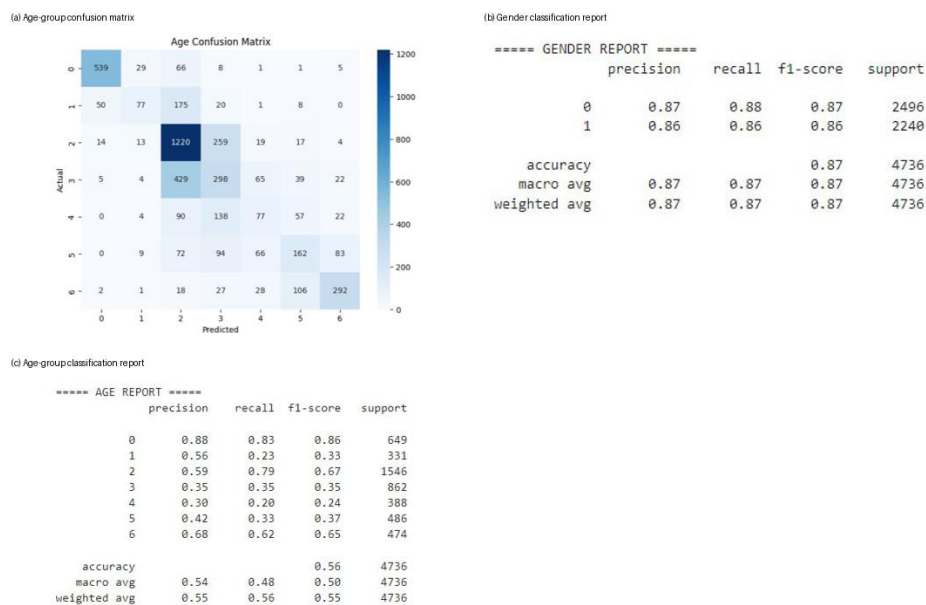


Figure 5. Downstream demographic evaluation associated with the restoration pipeline: (a) age-group confusion matrix, (b) gender-label classification report, and (c) age-group classification report.

These results support two conclusions. First, the restored-image workflow can feed a downstream attribute-analysis stage in the integrated application. Second, the age output is not sufficiently accurate to be treated as a precise forensic age estimate. The interface should therefore present age as an approximate category or investigative cue, accompanied by confidence and human review. The same caution applies to gender-related labels, particularly because dataset labels and appearance-based predictions do not represent a complete account of a person's identity.

6 | Discussion

6.1 | What the Results Demonstrate

The strongest evidence produced by the study is not that one model outperforms all existing forensic systems; such a claim would be unsupported because the modules use different data and operating conditions. The defensible result is that four heterogeneous analytical pathways were implemented behind a common case workflow and that each pathway can return a structured result to the user interface. The deepfake module shows strong within-distribution classification, the face module demonstrates small-scale identity integration, the SNP experiment shows that the selected feature representation is learnable by several classifiers, and the restoration path supports qualitative enhancement followed by measurable attribute classification.

The evaluation also reveals where the prototype should not overclaim. A 99.52% deepfake score does not imply robustness to unseen synthesis methods; a 94.55% five-identity result does not establish open-set face-search performance; the encoded SNP benchmark is not a substitute for independent forensic validation of published phenotype models; and age classification at 56% should be treated as approximate. Making these boundaries explicit strengthens rather than weakens the scientific argument because it keeps the contribution aligned with the evidence.

6.2 | Value of the Integrated Workflow

The case-centric design changes how the models are used. Instead of asking a practitioner to export an image to one tool, copy a result to a spreadsheet, send a second file to another tool, and manually rebuild the case context, the proposed platform keeps evidence selection and analysis invocation in one environment. Figure 6 illustrates this end-to-end path from the dashboard and active-case list to evidence upload, model selection, result generation, and case attachment. The second panel shows the four analytical interfaces exposed through the same mobile interaction pattern.

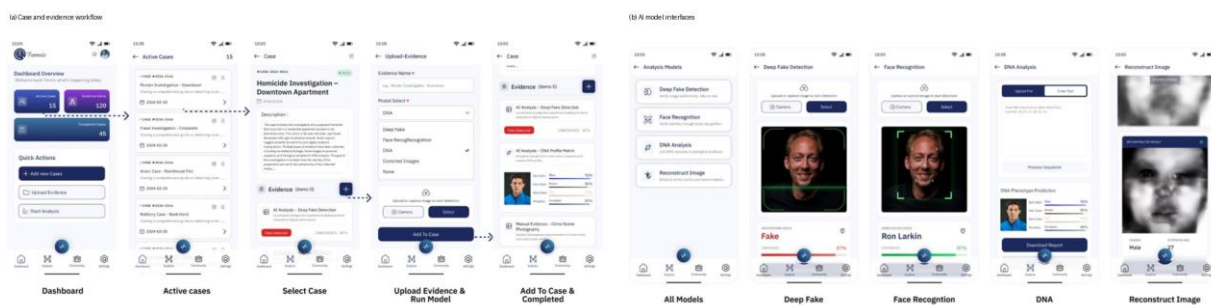


Figure 6. Case-centric operational workflow: (a) dashboard-to-case evidence handling and (b) the four analytical interfaces exposed through the same application.

The practical value is traceability at the application level: the user can understand which case is being examined, which evidence item is being analyzed, which model is invoked, and where the output is stored. This does not by itself establish forensic chain of custody, but it provides a coherent software foundation on which stronger provenance controls can be implemented.

6.3 | Human Review, Bias, and Evidentiary Boundaries

Forensic AI should support, not replace, professional reasoning. Clearview AI's own public guidance states that candidate facial-search results require independent human assessment, and NIST evaluations show that face-recognition performance varies across algorithms and populations [14, 17]. Similar caution is needed for DNA phenotype estimates and demographic classification. Probabilistic outputs can narrow an investigative search space, but they should not be presented as unique identification or as a legal conclusion.

The same principle applies to generative image restoration. GFP-GAN can generate visually plausible facial detail from degraded inputs [9], but plausible detail is not necessarily historically faithful detail. Restored images should therefore be clearly labeled as enhanced or reconstructed, and an original evidence file should remain available for examination. A restoration output should not silently replace source evidence or be treated as proof that generated fine detail existed in the original capture.

6.4 | Reliability and Reproducibility Considerations

At the system level, the modular architecture supports fault isolation and independent service updates. At the experimental level, reproducibility depends on preserving dataset partitions, preprocessing settings, model checkpoints, decision thresholds, software versions, and random seeds. The public components cited in this paper provide reference implementations or documented datasets, while the private development material

should be treated separately from the public test evidence. Future external validation should freeze these elements before testing to prevent inadvertent data leakage or post-hoc threshold selection.

7 | Illustrative Use Case and User Interaction

A typical professional session begins with authentication and entry to the dashboard. The practitioner opens an active case, reviews its evidence list, uploads or selects an evidence item, and chooses the relevant analytical service. For a suspicious image, the deepfake service can return an authenticity-screening result; for a facial comparison, the recognition service returns an identity candidate or similarity output; for genetic input, the DNA module presents phenotype-oriented results; and for a degraded face, the restoration service returns an enhanced image together with downstream demographic information. Figure 6 shows these steps directly in the interface.

The platform also includes a professional community and chatbot workflow, shown in Figure 7. The community area can support publication sharing and general knowledge exchange, while the chatbot can assist with navigation and general forensic questions. These features should be governed so that confidential case material is not inadvertently exposed outside the authorized case context. Their role is educational and collaborative; neither feature is part of the evidentiary decision path.

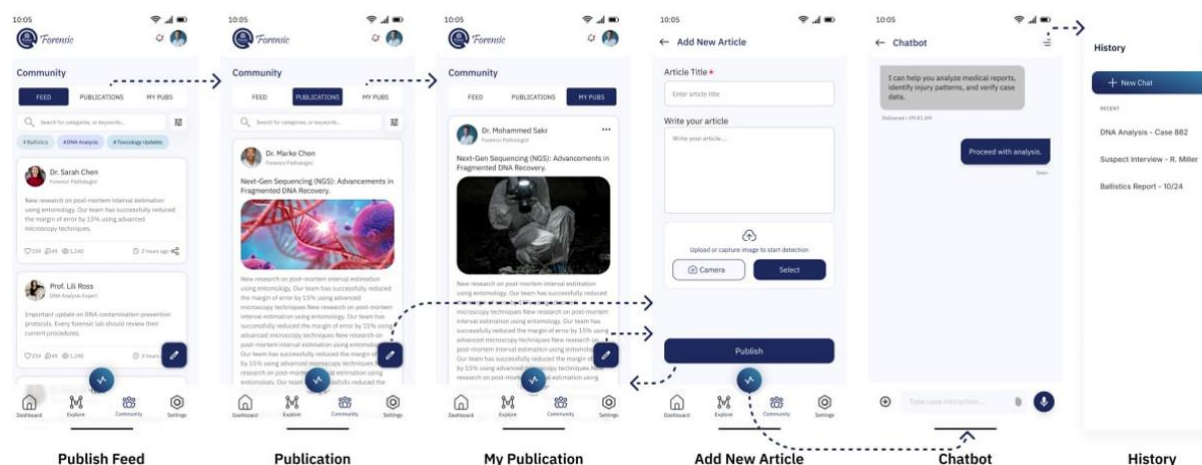


Figure 7. Community, publication, and chatbot workflow in the mobile interface.

8 | Practical Impact and Deployment Considerations

The platform is relevant to laboratories and investigative teams that need rapid access to several kinds of AI-assisted analysis without maintaining a separate user experience for each model. The architecture also supports research use because individual services can be replaced or benchmarked while the surrounding case workflow remains stable. This separation is especially useful when a laboratory wishes to compare alternative deepfake detectors, face embeddings, or phenotype classifiers without rebuilding case management.

For public users, the awareness and educational components can improve understanding of digital manipulation and forensic technologies. However, the professional and public roles should remain clearly separated because forensic evidence may contain biometric, genetic, and personally identifiable information. Production deployment should apply jurisdiction-appropriate privacy controls, retention policies, access auditing, and evidence-preservation procedures.

The research value of the system is therefore twofold: it provides a practical integration testbed for heterogeneous forensic AI services, and it exposes the methodological requirement that each service retain its own evaluation protocol and evidentiary limits. This is preferable to presenting a visually integrated application as if all of its modules had been validated to the same scientific standard.

9 | Conclusion and Future Work

This study presented an integrated AI-based forensic analysis platform for multimodal biometric, genetic, and digital evidence. Its principal contribution is the case-centric orchestration of deepfake screening, face recognition, SNP-based phenotype analysis, and face restoration within a common authenticated workflow. The architecture separates user access, case management, file handling, AI services, and persistence, allowing each analytical component to evolve without breaking the overall evidence workflow.

The evaluation demonstrates strong internal deepfake performance on a balanced 20,000-image test set, a successful but small five-identity face-recognition integration test, a high-performing Gradient Boosting model in the encoded SNP benchmark, and useful but uneven downstream demographic classification after the restoration path. Importantly, the paper distinguishes these results rather than combining them into one system accuracy, and it treats face restoration qualitatively because a reference-based restoration metric was not recorded.

Future work should prioritize external and cross-dataset validation, threshold-based biometric evaluation, explicit identity-preservation measurements for image restoration, calibrated uncertainty for DNA and demographic outputs, immutable evidence provenance, and formal security testing. These extensions would move the platform from a technically integrated prototype toward a validated forensic decision-support environment while preserving the central design principle established here: analytical automation should remain traceable to the evidence, interpretable in context, and subject to professional review.

Author Contribution

All authors contributed equally to this study.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, pp. 1-11, 2019. doi: 10.1109/ICCV.2019.00009.
- [2] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Canton Ferrer, "The DeepFake Detection Challenge (DFDC) Dataset," *arXiv:2006.07397*, 2020.
- [3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016.
- [4] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-stage Dense Face Localisation in the Wild," *arXiv:1905.00641*, 2019.
- [5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 4690-4699, 2019. doi: 10.1109/CVPR.2019.00482.
- [6] S. Walsh, F. Liu, K. N. Ballantyne, M. van Oven, O. Lao, and M. Kayser, "IrisPlex: A sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information," *Forensic Science International: Genetics*, vol. 5, no. 3, pp. 170-180, 2011. doi: 10.1016/j.fsigen.2010.02.004.
- [7] L. Chaitanya et al., "The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation," *Forensic Science International: Genetics*, vol. 35, pp. 123-135, 2018. doi: 10.1016/j.fsigen.2018.04.004.
- [8] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, "dbSNP: the NCBI database of genetic variation," *Nucleic Acids Research*, vol. 29, no. 1, pp. 308-311, 2001. doi: 10.1093/nar/29.1.308.

- [9] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards Real-World Blind Face Restoration with Generative Facial Prior," Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 9168-9178, 2021.
- [10] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data," Proc. IEEE/CVF Int. Conf. Computer Vision Workshops (ICCVW), pp. 1905-1914, 2021.
- [11] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 4401-4410, 2019. doi: 10.1109/CVPR.2019.00453.
- [12] Xhlulu, "140k Real and Fake Faces," Kaggle dataset, 70,000 real and 70,000 GAN-generated face images.
- [13] AICIP, "UTKFace: Large Scale Face Dataset," dataset repository, 20,000+ face images annotated by age, gender, and ethnicity.
- [14] P. Grother, M. Ngan, and K. Hanaoka, "Face Recognition Vendor Test Part 3: Demographic Effects," NISTIR 8280, National Institute of Standards and Technology, 2019. doi: 10.6028/NIST.IR.8280.
- [15] The Sleuth Kit, "Autopsy: Features," official product documentation.
- [16] Magnet Forensics, "Magnet AXIOM - Digital Forensic Software," official product documentation.
- [17] Clearview AI, "Facial Recognition for Investigations" and "Principles," official product documentation.
- [18] Parabon NanoLabs, "Snapshot DNA Phenotyping Service," official service documentation.

Disclaimer/Publisher's Note: The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.