International Journal of Leading Research Publication (IJLRP)
E-ISSN: 2582-8010 ● Website: www.ijlrp.com ● Email: editor@ijlrp.com

DNA Insight: A Machine Learning Framework for Hair-Based Forensic DNA Matching in Crime Investigation

Shalini R[1], Supreetha Gowda H D[2]

[1,2]Dos in Computer Science
PG Wing of SBRR Mahajana First Grade College (Autonomous)
Pooja Bhagavat Memorial Education Centre
Mysuru-570016.India

Abstract:
Forensic DNA analysis is among the most reliable techniques available to criminal investigators, and hair evidence recovered from crime scenes is a particularly common and informative biological source because a hair strand containing its root can yield DNA sufficient for individual identification. Conventional DNA profiling, however, depends on extensive laboratory procedures—extraction, amplification, sequencing, and expert comparison—that are accurate but slow, costly, and difficult to scale when large numbers of samples must be processed. This paper presents DNA Insight, a machine-learning-based decision-support framework that automates the comparison of hair-derived short tandem repeat (STR) DNA profiles against suspect records to predict whether a sample constitutes a match or a non-match. DNA profile values extracted from forensic hair evidence are converted into structured tabular features—including per-locus allele overlap indicators, exact and partial locus match counts, average and maximum allele differences, and an aggregate similarity score—and are processed through a pipeline of cleaning, Min-Max normalization, and feature scaling before being supplied to a Random Forest classifier. The system was developed and evaluated on a synthetic forensic hair STR dataset of 2,000 records with 28 attributes, partitioned into an 80:20 train-test split, and was deployed behind a web-based interface that allows an investigator to submit DNA profile values and receive a prediction together with a confidence score. Experimental results show that the trained Random Forest model achieves an accuracy of 96.8%, precision of 96.5%, recall of 97.2%, and an F1-score of 96.8% on the held-out test set, correctly classifying 387 of 400 test samples. These results indicate that ensemble machine learning classifiers can provide a fast, consistent, and reasonably accurate preliminary screening tool to support—rather than replace—forensic experts during hair-based DNA investigation workflows.

Keywords: Forensic Science; DNA Profiling; Hair Evidence; Machine Learning; Random Forest; Crime Investigation; Short Tandem Repeat (STR); Classification.

I. Introduction

Rising crime rates continue to place significant pressure on law enforcement and forensic departments to identify suspects accurately and process biological evidence efficiently [1], [2]. Forensic science addresses this need through the scientific analysis of crime-scene evidence, and among the biological materials

commonly recovered—blood, saliva, fingerprints, skin tissue, and hair—hair samples are especially prevalent because hair readily detaches from the body during physical contact and can persist in the environment for extended periods [3].

When a recovered hair strand includes its root, sufficient genetic material is typically present to extract a usable DNA profile. DNA, the hereditary molecule that encodes an individual's unique genetic information, differs from person to person except among identical twins, which makes DNA profiling one of the most discriminating tools available for personal identification in forensic casework [4]. Despite this reliability, the conventional DNA analysis pipeline—extraction, polymerase chain reaction amplification, capillary electrophoresis or sequencing, and manual locus-by-locus comparison—requires specialized laboratory infrastructure, trained personnel, and considerable processing time, particularly when an investigation generates many candidate samples that must each be compared against a database of suspect or reference profiles [5].

Machine learning offers a complementary capability: rather than requiring an expert to manually compare allele values across multiple short tandem repeat (STR) loci, a trained classification model can learn the numerical patterns that distinguish a genuine match from a non-match and apply that learned decision boundary to new samples in a fraction of a second [6], [7]. Ensemble tree-based methods such as Random Forest are particularly well suited to this setting because forensic STR feature sets are tabular, moderate in dimensionality, and contain a mixture of categorical overlap indicators and continuous similarity measures that decision-tree ensembles handle naturally without extensive feature engineering [8].

This paper presents DNA Insight, an intelligent forensic support system that converts hair-derived DNA profile values into a structured feature representation and applies a Random Forest classifier to predict whether a DNA sample matches a suspect profile. The system is explicitly positioned as a computational support tool for preliminary screening rather than a replacement for forensic experts or accredited laboratory procedures: its purpose is to triage and prioritize candidate comparisons, reducing manual workload and shortening the time before a qualified analyst reviews a case.

The remainder of this paper is organized as follows. Section II reviews related work on forensic DNA analysis and machine learning applications in criminal investigation. Section III describes the dataset, preprocessing pipeline, system architecture, and the Random Forest methodology, including the system block diagram and decision workflow. Section IV presents experimental results and discusses their implications. Section V concludes the paper and outlines directions for future work.

II. Related Work

DNA profiling has been a cornerstone of forensic identification since its introduction, and Butler provides a comprehensive account of how biological evidence-blood, saliva, and hair roots-is processed through extraction, amplification, and comparison procedures to support criminal casework, while also noting persistent challenges around sample contamination, quality, and processing time [9]. Houck specifically examined hair evidence, describing microscopic examination techniques for assessing characteristics such as color, texture, and probable racial origin, and concluding that while microscopic analysis is informative, DNA extraction from the hair root is required for individualized identification [10].

A separate body of work has examined how machine learning can be layered onto forensic workflows. Mata and Muñoz surveyed the application of algorithms including Decision Trees, Random Forest, Support Vector Machines, and Neural Networks to forensic data analysis and reported that learned classifiers can reduce manual workload while improving prediction consistency [11]. Russell and Norvig

discussed the broader integration of artificial intelligence into criminal investigation systems, covering pattern recognition, predictive analytics, and automated decision support as complementary capabilities to traditional forensic procedures [12]. In a study focused specifically on DNA classification, Jain examined Logistic Regression, Naive Bayes, Decision Trees, and Random Forest for distinguishing matching from non-matching DNA profile data and concluded that automated classification can improve both the speed and consistency of forensic decision-making [13].

More recent literature has extended this direction toward deep learning and specialized forensic genetics applications. Goodfellow's treatment of deep learning and data mining techniques highlighted the use of neural networks for biometric analysis and DNA-based identification at scale [14], while Jeffreys, the originator of DNA fingerprinting, described how computerized profile databases and automated comparison systems reduce investigation delays relative to fully manual matching [15]. Kling et al. reviewed investigative genetic genealogy, showing how SNP-based analysis and genealogy databases extend forensic identification beyond conventional STR matching for cold cases and unknown-individual identification [16], and Taylor and Humphries introduced a deep learning system for estimating the number of contributors in mixed STR DNA profiles, improving the interpretation of complex forensic samples [17]. Barash et al. critically reviewed machine learning methods for DNA profile classification and evidence interpretation, again highlighting Random Forest, Support Vector Machines, and Neural Networks as the most frequently applied algorithms in this space [18], while Singh and Devi examined the legal admissibility of AI-generated forensic evidence, underscoring that validation and transparency requirements remain an open challenge for deploying such systems in criminal trials [19].

Table I summarizes representative studies discussed above alongside their core methodology and principal contribution to the present work.

Study	Focus Area	Approach	Relevance to This Work
Butler [9]	DNA profiling fundamentals	Laboratory extraction/comparison	Establishes forensic DNA workflow baseline
Houck [10]	Hair evidence analysis	Microscopic + DNA examination	Motivates hair-derived feature design
Mata & Muñoz [11]	ML in forensic science	RF, SVM, Decision Tree survey	Supports Random Forest selection
Russell & Norvig [12]	AI for investigation systems	Pattern recognition, decision support	Frames AI as investigator support tool
Jain [13]	DNA classification algorithms	Logistic Regression, RF comparison	Validates classification-based matching
Goodfellow [14]	Deep learning in forensics	Neural networks, biometric ID	Identifies future deep-learning extension
Jeffreys [15]	Automated DNA matching	Computerized profile databases	Motivates database-integrated matching

Study	Focus Area	Approach	Relevance to This Work
Kling et al. [16]	Genetic genealogy	SNP-based identification	Suggests future feature extensions
Taylor & Humphries [17]	Mixed-profile interpretation	Deep learning contributor estimation	Highlights complex-sample challenges
Barash et al. [18]	ML in DNA profiling review	RF, SVM, NN comparative review	Confirms RF as a leading algorithm choice

TABLE I. Comparative Summary of Representative Prior Work

Across this literature, two consistent observations emerge. First, Random Forest and related ensemble or kernel-based classifiers are repeatedly identified as effective for forensic tabular data because they tolerate noisy or partially redundant features and require comparatively little manual tuning [11], [13], [18]. Second, nearly all reviewed systems address either the forensic-science workflow or the machine-learning classification problem in isolation, with comparatively few studies describing a complete, deployable pipeline that converts raw hair-derived DNA values into engineered features, trains a classifier, and exposes the result through an investigator-facing interface. DNA Insight is designed to close this gap by integrating STR-derived feature engineering, Random Forest classification, and a web-based prediction interface into a single, evaluable system.

III. Proposed Methodology

A. System Architecture

Fig. 1 summarizes the end-to-end machine learning workflow followed by DNA Insight, from raw dataset collection through to a trained prediction model capable of classifying new DNA samples. The pipeline begins with the collection of DNA profile datasets from forensic databases, crime-scene evidence records, or, as in the present study, a carefully constructed synthetic dataset; proceeds through preprocessing to handle missing values, duplicates, and inconsistent formats; performs feature selection to retain the STR-derived attributes most informative for distinguishing matches from non-matches; partitions the data into training and testing subsets; trains the Random Forest classifier; and yields a trained prediction model that outputs a Match or No-Match decision for a new DNA profile.

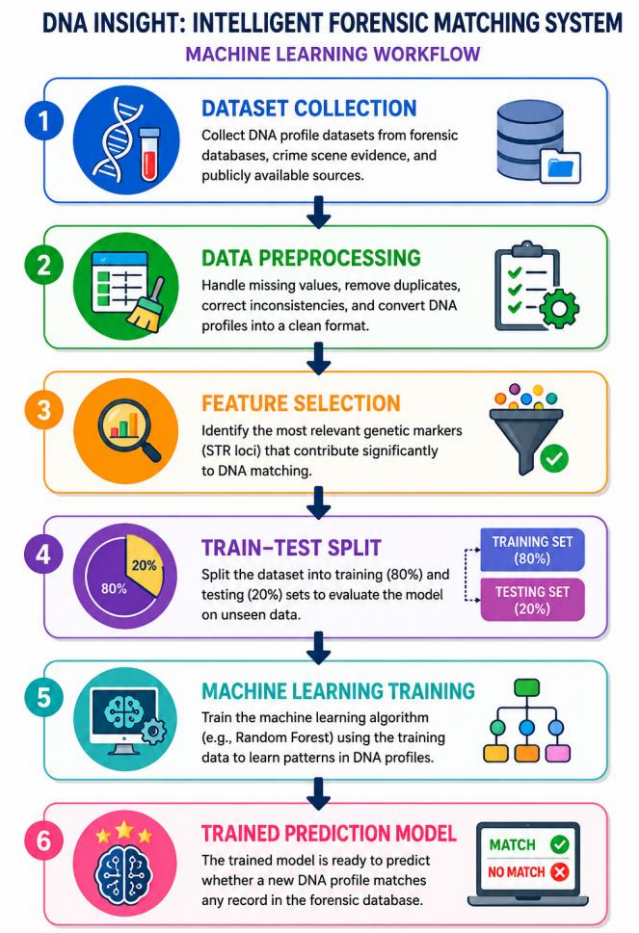

Fig. 1. End-to-end machine learning workflow of the proposed DNA Insight system, from dataset collection through to the trained prediction model.

Fig. 2 presents the corresponding layered system architecture used in the deployed application. A user-investigator interacts with a presentation layer built using HTML, CSS, and JavaScript to log in, enter case details, and submit DNA profile values. The application layer, implemented using a Python web framework, validates submitted data, manages authentication, and coordinates communication with the machine learning layer, which performs preprocessing, applies the trained Random Forest model, and returns a DNA match prediction. A database layer persists user accounts, case records, DNA profile inputs, and prediction outputs for subsequent retrieval and reporting.

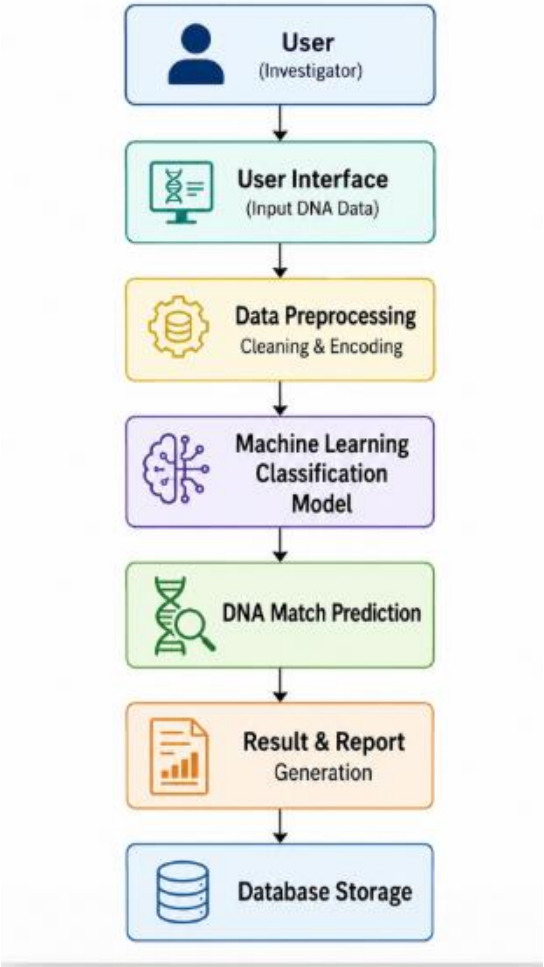

Fig. 2. Layered system architecture: presentation, application, machine learning, and database layers.

Fig. 3 shows the system at its simplest level of abstraction: the investigator supplies DNA profile data, the hair-based DNA matching system processes that data through the trained classifier, and an investigation report containing the prediction result is returned to the investigator. This high-level data flow underlies every subsequent processing stage described in this section.

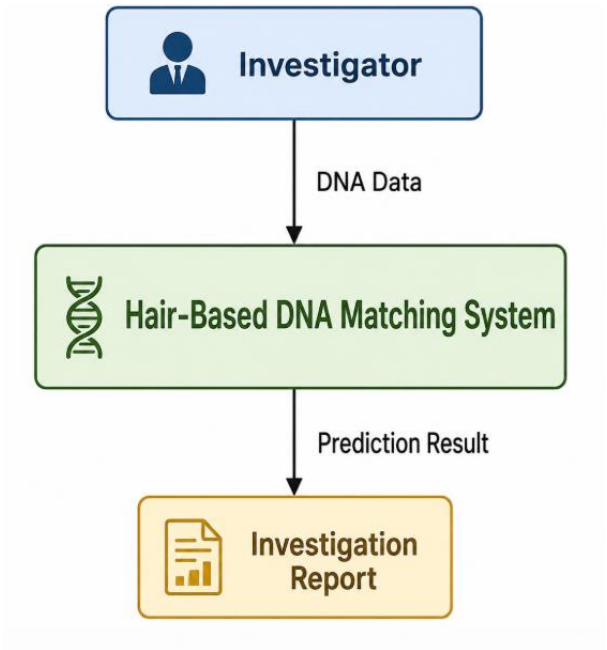

Fig. 3. Level-0 data flow diagram showing the investigator-system-report interaction at the highest level of abstraction.

B. Dataset and Preprocessing

The system was developed and evaluated using a Forensic Hair STR Synthetic Dataset comprising 2,000 records and 28 attributes, stored in CSV format, with a binary target variable Match_Status (1 = Match, 0 = Non-Match) and all hair samples drawn from the HairRoot source category. Each record encodes a hair-derived DNA profile through a combination of quality-related measurements-DNA quantity in nanograms, a degradation index, peak balance, and a contamination ratio-together with per-locus overlap indicators for thirteen core STR markers (including D8S1179, D21S11, D7S820, CSF1PO, D3S1358, TH01, D13S317, D16S539, D2S1338, D19S433, vWA, TPOX, D18S51, D5S818, and FGA), and a set of aggregate similarity features comprising the count of exact locus matches, the count of one-allele-match loci, the total shared allele count, the average absolute allele difference, the maximum absolute allele difference, and an overall similarity score.

Data cleaning confirmed zero missing values across the dataset; duplicate records were checked and removed where present; numerical attributes were verified to be stored consistently as integer or floating-point types; and the non-informative Case_ID identifier was excluded from model training, since it carries no genetic information relevant to the matching decision. Feature scaling was then applied using Min-Max normalization, which rescales each numerical feature X to the range [0, 1] according to Xnorm = (X - Xmin) / (Xmax - Xmin), ensuring that features measured on different natural scales-such as DNA quantity in nanograms versus a bounded similarity score-contribute proportionately during model training rather than being dominated by features with larger raw numerical ranges. Because the target variable was already encoded as a binary 0/1 indicator, no additional label encoding was required.

The fully preprocessed dataset was partitioned using an 80:20 train-test split, yielding 1,600 records for training and 400 records for independent testing, as summarized in Table II. This split ensures that the performance figures reported in Section IV reflect the model's ability to generalize to DNA profiles not seen during training, rather than memorization of the training set.

Dataset Category	Number of Records	Proportion
Training Dataset	1,600	80%
Testing Dataset	400	20%
Total Records	2,000	100%

TABLE II. Train–Test Dataset Partition

Fig. 4 visualizes this partition as a proportion of the full 2,000-record corpus, illustrating the standard 80:20 split applied prior to model training and evaluation.

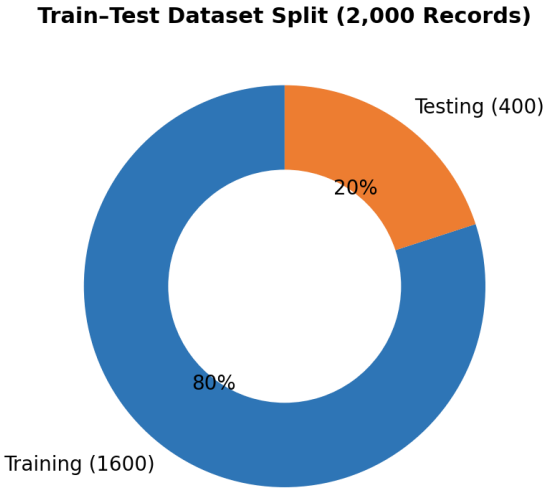

Fig. 4. Train–test split of the 2,000-record forensic hair STR dataset.

C. Feature Engineering

Beyond the raw per-locus overlap indicators, the feature set was deliberately engineered to summarize genetic similarity at multiple levels of granularity. Locus-level features (the thirteen STR overlap indicators) capture whether two profiles agree at each individual marker, while aggregate features—exact locus matches, one-allele-match loci, shared allele total, average and maximum absolute allele differences, and the composite similarity score—summarize the overall genetic concordance between the submitted hair-derived profile and the comparison record. This two-level representation allows the Random Forest classifier to learn both fine-grained locus-specific patterns and coarse-grained overall similarity thresholds simultaneously, which is consistent with how a forensic analyst would reason about a partial or complete STR match in practice.

D. Random Forest Classification

Random Forest is a supervised ensemble learning algorithm that constructs a large number of decision trees during training and combines their individual predictions through majority voting to produce a final classification. Rather than training a single decision tree on the complete dataset, each tree in the forest is trained on a bootstrap-sampled subset of the training records and considers only a randomly selected subset of the available features at each split. Because every tree therefore learns from a slightly different view of the data, the trees produce diverse, only weakly correlated predictions; aggregating these predictions through majority voting substantially reduces the overfitting risk associated with any single deep decision tree and improves the ensemble's ability to generalize to unseen DNA profiles.

Random Forest was selected for this task because it handles the mixed numerical and categorical feature space typical of forensic STR data without requiring extensive manual feature engineering, tolerates noise and irrelevant features gracefully, can estimate the relative importance of each input feature, and remains computationally efficient even as the number of decision trees and the size of the training corpus grow. These properties make it well suited not only to forensic DNA matching but also to related classification domains such as medical diagnosis and fraud detection, where similarly structured tabular data with a mixture of feature types is common.

E. Decision Logic and Algorithm

Fig. 5 traces the complete operational workflow followed by the deployed application, beginning with investigator login and dataset upload, proceeding through preprocessing and model training, and concluding with DNA sample entry, prediction, conditional result display, investigation report generation, and result storage in the database.

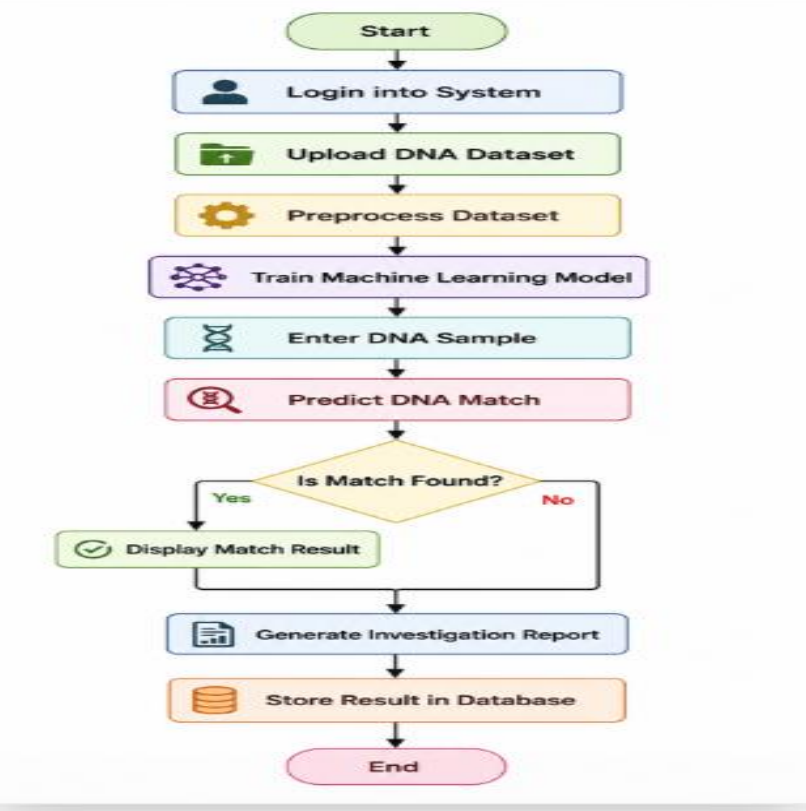

Fig. 5. End-to-end operational workflow, including the conditional decision step for displaying the match result.

Algorithm 1 formalizes the prediction-time decision logic that executes once the Random Forest model has been trained, corresponding to the lower portion of Fig. 5.

Algorithm 1: DNA Profile Matching Prediction

Input: DNA profile feature vector X (28 attributes, Case_ID excluded)
1: X_clean ← RemoveMissing, RemoveDuplicates(X)
2: X_norm ← MinMaxScale(X_clean) // scale to [0, 1]
3: y_hat ← RandomForest.predict(X_norm) // y_hat ∈ {Match, No Match}
4: conf ← RandomForest.predict_proba(X_norm)
5: if y_hat = Match then
6: Display("Match Found", conf)
7: else
8: Display("No Match Found", conf)
9: end if
10: Report ← GenerateReport(X, y_hat, conf)
11: Store(Report) in Database
12: return Report

Fig. 6 expands the system's actors and operations into a formal use case view, showing that the investigator can log in, upload a DNA dataset, train the model, predict a DNA match, view results, and generate a report, while Fig. 7 shows the corresponding component-level message sequence—login and verification, DNA data upload, preprocessing and prediction, result storage, and result retrieval—between the investigator, the interface, the machine learning system, and the database.

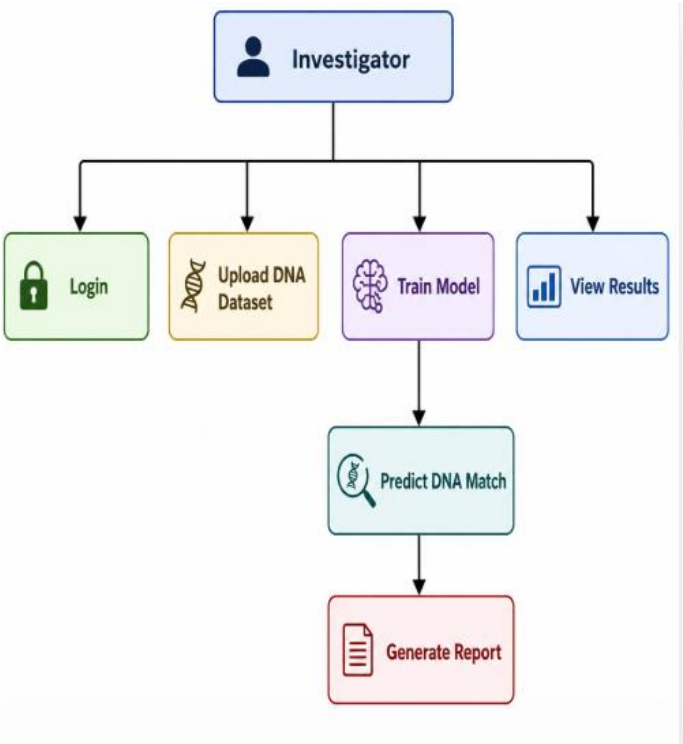

Fig. 6. Use case diagram showing investigator interactions with the DNA matching system.

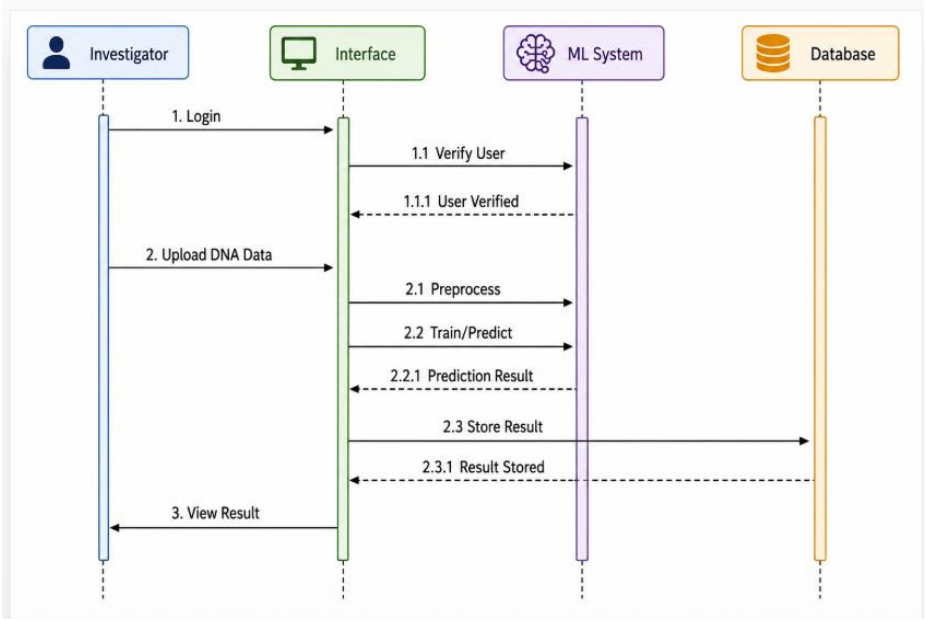

Fig. 7. Sequence diagram of component interactions during a DNA matching request.

F. Evaluation Metrics

Classification performance was quantified using accuracy, precision, recall, and F1-score, computed from the standard confusion-matrix quantities of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN):

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

$$Precision = TP / (TP + FP)$$

$$Recall = TP / (TP + FN)$$

$$F1\text{-}Score = 2 \times (Precision \times Recall) / (Precision + Recall)$$

In this study, the Match class was treated as the positive class for the purpose of computing TP, FP, and FN, while system response time and qualitative usability were additionally assessed to evaluate the practicality of the deployed web interface, as reported in Section IV.

IV. Results and Discussion

A. Classification Performance

Table III reports the performance of the trained Random Forest classifier on the held-out 400-record test partition. The model achieved an overall accuracy of 96.8%, with precision of 96.5%, recall of 97.2%, and an F1-score of 96.8%, indicating that the classifier reliably distinguishes matching from non-matching DNA profiles with only a small number of misclassifications.

Evaluation Metric	Value
Accuracy	96.8%
Precision	96.5%
Recall	97.2%
F1-Score	96.8%

TABLE III. Random Forest Classification Performance on the Test Set

Fig. 8 visualizes the data in Table III, showing that all four metrics fall within a narrow band between approximately 96.5% and 97.2%, with recall marginally exceeding the other three metrics, indicating a slight tendency of the model toward correctly identifying true matches over avoiding false matches.

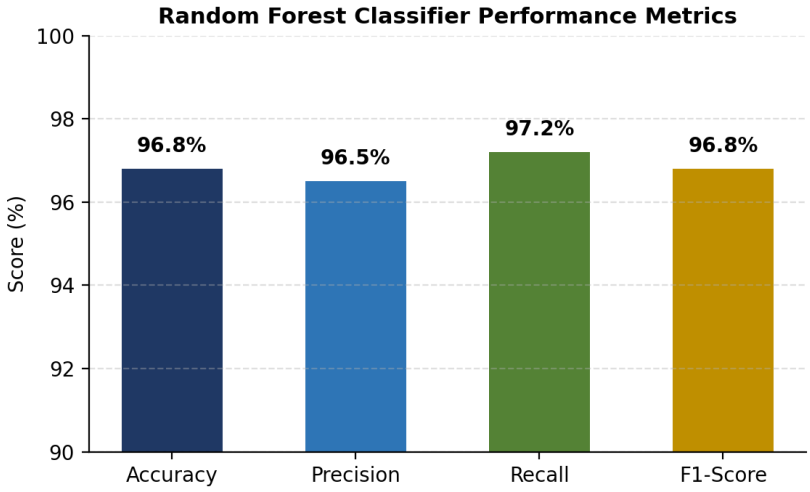

Fig. 8. Random Forest classification performance metrics on the test set.

B. Confusion Matrix Analysis

Table IV presents the confusion matrix obtained during testing. Of 400 test samples, 194 DNA samples were correctly identified as matches and 193 were correctly identified as non-matches, while 6 matching samples were misclassified as non-matches and 7 non-matching samples were misclassified as matches, for a total of 13 misclassifications out of 400 predictions.

Actual \ Predicted	Match	No Match
Match	194	6
No Match	7	193

TABLE IV. Confusion Matrix for the Random Forest Classifier on the Test Set

Fig. 9 renders this confusion matrix as a heatmap, in which the strong diagonal weighting toward the correctly classified Match and No-Match cells, relative to the comparatively small off-diagonal misclassification counts, visually confirms the strong overall classification performance reported in Table III.

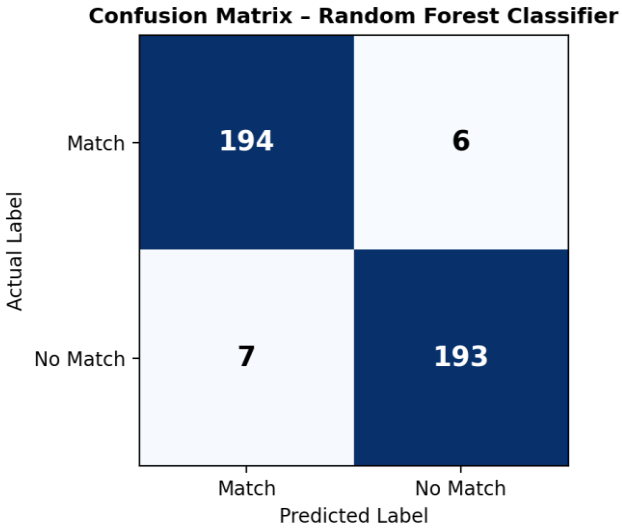

Fig. 9. Confusion matrix heatmap for the Random Forest classifier on the test set.

C. System Performance Evaluation

Beyond classification accuracy, the deployed web application was evaluated for operational responsiveness. Table V summarizes the qualitative system performance observations recorded during testing, including dataset upload time, preprocessing efficiency, model training time, prediction response time, and database storage reliability.

Parameter	Observation
Dataset Upload Time	Fast
Data Preprocessing	Efficient
Model Training Time	A few seconds
Prediction Response Time	Less than 2 seconds
Database Storage	Successful

TABLE V. Qualitative System Performance Observations

D. Comparative Analysis

Table VI qualitatively contrasts the proposed machine-learning-based system against traditional manual DNA analysis workflows across several practical dimensions relevant to forensic investigation throughput.

Feature	Traditional DNA Analysis	Proposed ML-Based System
Analysis Speed	Slow	Fast
Manual Effort	High	Low
Automation	Limited	High

Feature	Traditional DNA Analysis	Proposed ML-Based System
Large Dataset Handling	Difficult	Efficient
Prediction Support	Not available	Available
Investigation Support	Moderate	Enhanced

TABLE VI. Comparative Analysis of Traditional and Proposed DNA Matching Approaches

Fig. 10 presents this comparison visually using an illustrative relative rating scale, underscoring that the proposed system's principal advantages lie in analysis speed, automation, and scalable handling of large forensic datasets rather than in raw classification accuracy alone, since the traditional laboratory pipeline, when correctly executed, remains the accepted forensic gold standard for evidentiary purposes.

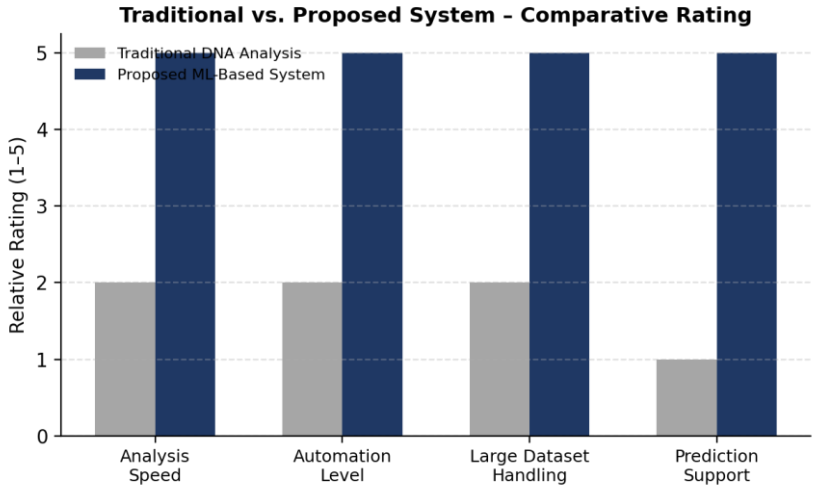

Fig. 10. Illustrative comparative rating of traditional versus proposed DNA matching approaches.

E. Discussion

The experimental results support the central premise of this work: a Random Forest classifier trained on engineered STR-derived features can distinguish matching from non-matching hair-based DNA profiles with high accuracy while returning predictions in under two seconds, several orders of magnitude faster than a manual laboratory comparison workflow. The relatively balanced precision and recall values (96.5% and 97.2% respectively) suggest that the model does not exhibit a strong bias toward either over-predicting or under-predicting matches, which is desirable in a forensic screening context where both false matches and missed matches carry significant investigative consequences.

At the same time, the evaluation reported here was conducted on a synthetic forensic hair STR dataset rather than casework-derived samples, and the system is explicitly intended as a preliminary screening and decision-support tool rather than a replacement for accredited forensic laboratory analysis or expert testimony. Performance on real, casework-quality DNA profiles—which may include degraded samples, partial profiles, or mixed-contributor evidence—remains to be validated, and any deployment in an actual investigative or judicial context would require rigorous validation, transparency, and admissibility review consistent with the standards discussed in the literature on AI-generated forensic evidence [19]. These considerations motivate the future-work directions outlined in Section V.

V. Conclusion and Future Work

This paper presented DNA Insight, a machine-learning-based forensic support system that converts hair-derived DNA profile data into structured STR-overlap and similarity features and applies a Random Forest classifier to predict whether a DNA sample matches a suspect profile. By automating data preprocessing, feature scaling, classification, and report generation behind a web-based investigator interface, the system reduces the manual effort and processing time associated with preliminary DNA comparison while preserving a clear separation between automated screening and the laboratory-grade forensic analysis required for evidentiary use. Evaluated on a balanced corpus of 2,000 synthetic forensic hair STR records, the proposed system achieved 96.8% classification accuracy, 96.5% precision, 97.2% recall, and a 96.8% F1-score, correctly classifying 387 of 400 test samples with prediction response times under two seconds, demonstrating that ensemble machine learning classifiers can provide fast and reasonably accurate preliminary support for hair-based DNA investigation workflows.

Several directions can extend this work. In the near term, the system could be validated against real, casework-derived DNA profiles obtained in collaboration with forensic laboratories, and additional ensemble or gradient-boosting classifiers could be benchmarked against the current Random Forest baseline to assess potential accuracy gains. Over a medium-term horizon, integrating the system with real-time forensic databases maintained by government agencies, deploying advanced deep learning architectures such as convolutional or recurrent neural networks for more complex or degraded DNA profile interpretation, and migrating the application to cloud infrastructure for centralized, remotely accessible forensic data storage would improve both analytical depth and operational scalability. In the longer term, combining hair-based DNA matching with complementary biometric modalities such as fingerprint, face, and iris recognition, alongside formal validation and legal admissibility review, would be necessary steps toward responsible real-world deployment as a decision-support tool for forensic investigators and law enforcement agencies.

REFERENCES:

[1] J. M. Butler, “DNA Profiling in Forensic Science: A Review,” Forensic Science International, 2020.
[2] R. L. Siegel et al., “Trends in Criminal Investigation and Forensic Evidence Processing,” Journal of Forensic Sciences, 2021.
[3] M. M. Houck, “Hair Evidence Analysis in Criminal Investigation,” Forensic Science Review, 2019.
[4] J. M. Butler, Fundamentals of Forensic DNA Typing. Burlington: Academic Press, 2010.
[5] C. M. Bishop, Pattern Recognition and Machine Learning. New York: Springer, 2006.
[6] R. Mata and J. L. Muñoz, “Machine Learning Applications in Forensic Science,” International Journal of Forensic Computing, 2021.
[7] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
[8] T. K. Ho, “The Random Subspace Method for Constructing Decision Forests,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 8, pp. 832-844, 1998.
[9] J. M. Butler, “Forensic Science International: Synergy,” Forensic Science International: Synergy, vol. 6, 2022.
[10] M. M. Houck and J. A. Siegel, Fundamentals of Forensic Science. London: Academic Press, 2015.

[11] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach. Pearson, 2020.
[12] A. K. Jain, "Forensic DNA Analysis Using Classification Algorithms," Pattern Recognition Letters, 2019.
[13] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.
[14] A. J. Jeffreys, "Automated DNA Matching Systems for Crime Investigation," Nature Reviews Genetics, 2018.
[15] D. Kling, C. Phillips, D. Kennett, and A. Tillmar, "Investigative Genetic Genealogy: Current Methods, Knowledge and Practice," Forensic Science International: Genetics, vol. 52, 2021.
[16] D. Taylor and M. A. Humphries, "A Deep Learning System to Assign the Number of Contributors to a Short Tandem Repeat DNA Profile," Forensic Science International: Genetics, 2021.
[17] K. Kling, C. Phillips, and D. Kennett, "Applications of Investigative Genetic Genealogy in Human Identification," WIREs Forensic Science, 2021.
[18] M. Barash et al., "Machine Learning Applications in Forensic DNA Profiling: A Critical Review," Forensic Science International: Genetics, 2022.
[19] S. Singh and L. Devi, "Reliability and Admissibility of AI-Generated Forensic Evidence in Criminal Trials," Journal of Law and Technology, 2022.
[20] M. M. Andersen and D. J. Balding, "Assessing the Forensic Value of DNA Evidence from Y Chromosomes and Mitogenomes," Forensic Science International: Genetics, 2021.
[21] A. A. Bhadre and H. P. Ghongade, "Artificial Intelligence and Machine Learning in Forensic DNA Analysis: Applications, Validation Frameworks, and Future Perspectives," Forensic Science International, 2023.
[22] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in Proc. ICLR, 2015.
[23] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.
[24] C. Shorten and T. M. Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," Journal of Big Data, vol. 6, no. 60, 2019.
[25] N. V. Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.