Original Research Article
ISSN: 3041-5403
Journal Homepage: https://bitsjournal.researchfloor.org/

A Comparative Evaluation of Machine Learning and Ensemble Models for Predicting Digital Forensic Investigation Outcomes Using Network Intrusion Data

Bashirat Aderayo Bamigboye

National Open University of Nigeria (NOUN), Abuja Nigeria

Citation: Bashirat Aderayo Bamigboye (2026). A Comparative Evaluation of Machine Learning and Ensemble Models for Predicting Digital Forensic Investigation Outcomes Using Network Intrusion Data. *Journal of Business, IT, and Social Science*.
DOI: https://doi.org/10.51470/BITS.2026.05.01.38

Corresponding Author: **Bashirat Aderayo Bamigboye** | E-Mail: **(Bbashirat4joy@gmail.com)**

07 November 2025: Received | 03 December 2025: Revised | 09 January 2026: Accepted | 01 February 2026: Available Online

ABSTRACT

The increasing complexity of cybercrime and the growing volume of digital evidence have intensified the need for accurate and reliable predictive models in Forensic Science. Traditional statistical approaches often struggle to handle high-dimensional and heterogeneous network data, thereby limiting their effectiveness in modern digital forensic investigations. This study evaluates the performance of machine learning and ensemble techniques in predicting digital forensic investigation outcomes using intrusion detection datasets. A quantitative experimental research design was adopted using CIC-IDS2017, UNSW-NB15, and NSL-KDD. Logistic Regression was employed as the baseline model, while Random Forest, Support Vector Machine, and Gradient Boosting were implemented as advanced predictive models. Data preprocessing included normalisation, feature selection, and the removal of data leakage to ensure validity. Model performance was evaluated using accuracy, precision, recall, F1 score, Area Under the Receiver Operating Characteristic Curve, and error-based metrics. The results indicate that machine learning models significantly outperform the baseline statistical approach. Logistic Regression achieved an accuracy of 84.2 percent, while Random Forest and Support Vector Machine exceeded 92 percent. Gradient Boosting achieved the highest accuracy of 95.2 percent with the lowest false positive and false negative rates. The findings further reveal that data preprocessing plays a critical role in ensuring reliable results, as initial inflated performance was linked to data leakage. The study concludes that ensemble learning techniques provide superior predictive performance for digital forensic investigations. However, effective deployment requires a careful balance between predictive accuracy and interpretability.

Keywords: *Machine Learning, Digital Forensic Investigation, Network Intrusion Detection, Ensemble Models, Predictive Modelling.*

INTRODUCTION

Background to the Study

The field of modern Forensic Science investigations has been substantially transformed by the rapid increase in cybercrime and the increasing volume of digital evidence. The demand for analytical techniques that can handle large, complex, and high-dimensional data that originates from digital settings is on the rise. Forensic analysis underwent a significant transformation from 2019 to 2025, transitioning from expert-based judgement to data-driven predictive modelling. This improved the process's consistency, efficiency, and accuracy [13]. Expert analysis and statistical techniques, including logistic regression and likelihood ratio models, were employed in forensic investigations. These methods offer systematic approaches for deriving inferences; however, they are ineffective for modelling nonlinear relationships or managing high-dimensional datasets. The limitations have become more apparent as the complexity of cybercrime has increased and the prevalence of network-based evidence has increased [9]. In order to confront these obstacles, a multitude of individuals have implemented supervised machine learning methodologies, including Random Forest, Support Vector Machine, and Gradient Boosting.

These models are capable of identifying intricate feature interactions and nonlinear patterns, which leads to enhanced forecast dependability and classification accuracy. In classification and anomaly detection tasks, machine learning techniques have been demonstrated to outperform traditional statistical methods in numerous empirical studies. Nevertheless, their efficacy is significantly influenced by data quality, feature engineering, and class imbalance management [8]. The integration of deep learning has improved forensic analytics by automating feature extraction and enhancing pattern identification. Nevertheless, challenges persist, particularly in digital forensic contexts that involve network intrusion data, despite these advantages. Significant challenges include traffic fluctuation, evolving attack patterns, and generalisation across datasets [5].

The predictive accuracy and data processing capabilities of a model are both critical factors in determining its efficacy. Noise, absent values, and class imbalance are frequently present in digital forensic datasets, which can lead to bias and distortion of results. In forensic situations, such distortions may result in the inability to identify malevolent conduct or incorrect suspicion, thereby jeopardising the integrity of the evidence [10]. The comprehensibility and applicability of the information in a judicial context are also significant concerns.

Forensic evidence is required to adhere to legal standards of scientific validity, repeatability, and transparency. However, a significant number of high-performing models are operated as intricate systems with restricted interpretability. This has led to a surge in interest in AI methodologies that are easier to understand and that facilitate the legal adoption of AI [7] [12]. Moreover, ensemble learning techniques such as stacking and boosting have exhibited superior performance and improved system stability. The comprehension and substantiation of forensic applications are further complicated by the additional complexity that these models introduce. Unfortunately, the majority of research conducted to date has concentrated on biometrics and DNA analysis, despite these advancements. Limited research has been conducted on digital forensic environments that employ network intrusion data. There is a dearth of comprehensive studies that assess prediction accuracy, data reliance, interpretability, and risk-sensitive decision-making within a unified framework. An integrating explainability and risk-based assessment, this study evaluates a variety of machine learning techniques using network intrusion datasets to address this disparity.

Statement of the Problem

The capacity of forensic investigators to employ analytical techniques has been considerably impeded by the rapid escalation of cybercrime and the growing volume and complexity of digital data. Conventional methods, including logistic regression and a variety of statistical models, offer structured and comprehensible methods for data analysis. Nevertheless, they are unable to fully capture the complex nonlinear interactions that are common in high-dimensional network intrusion data. In contrast, machine learning techniques, such as ensemble models, have demonstrated enhanced prediction effectiveness in the detection and classification of detrimental actions. In spite of these developments, their applicability in forensic contexts is still restricted by concerns regarding comprehensibility, reliability, and admissibility in court [9].

The inherent trade-off between prediction accuracy and interpretability is a significant issue in this domain. High-performing models that frequently operate as intricate systems with minimal transparency are exemplified by Gradient Boosting and other ensemble methods. This complicates the elucidation of their decision-making processes in accordance with legal and forensic standards, statistical models that are more straightforward are more comprehensible; however, they demonstrate a decrease in predictive accuracy. This trade-off is a substantial obstacle in forensic research, as analytical procedures must be legally defensible, reproducible, transparent, and accurate [7] [12]. The dependability and efficacy of machine learning models are considerably impacted by the quality of the data, which can be compromised by factors such as class imbalance, noise, and feature selection. These factors have the potential to substantially impact model outcomes, resulting in false positives or false negatives. These outcomes can have severe repercussions for forensic investigations. The implications of these errors in a risk-sensitive forensic setting are frequently disregarded in current research, which predominantly evaluates models based on accuracy measures [13]. Despite the fact that previous research has investigated the application of machine learning methods in cybersecurity and forensic science, there is a dearth of comprehensive empirical studies that assess the predictive performance and interpretability of digital forensic investigation outcomes, particularly when applied to network intrusion datasets. The critical necessity for analytical models that are balanced, reliable, and legally admissible is largely overlooked in recent research, which primarily concentrates on either performance optimisation or interpretability separately. The primary issue that this paper addresses is the lack of a systematic and unified evaluation of machine learning and ensemble techniques that effectively balance predicted accuracy with interpretability in digital forensic investigations. It is essential to resolve this issue to create analytical models that are technically robust and comply with the evidentiary standards necessary in legal and forensic contexts. This study aims to address this lacuna by conducting a comparative evaluation of statistical, machine learning, and ensemble models within a unified framework that incorporates both performance and interpretability considerations.

Research Objectives

The main objective of this study is to comparatively evaluate machine learning and ensemble models for predicting digital forensic investigation outcomes using network intrusion data. While the specific objectives are:

1. To evaluate and compare the predictive performance of statistical, supervised machine learning, and ensemble models in forecasting digital forensic investigation outcomes using network intrusion datasets.
2. To assess the effect of data preprocessing techniques, including feature selection, normalization, and handling of class imbalance, on the performance and generalizability of the models.
3. To analyze model performance using both accuracy-based and error-based evaluation metrics, including false positive rate and false negative rate.
4. To implement a cost-sensitive evaluation framework to quantify forensic risk associated with misclassification errors, and determine the most reliable model for forensic decision-making.
5. To compare the effectiveness of ensemble learning methods with single algorithm models in terms of predictive accuracy, robustness, and risk minimisation.

RELATED WORK

In forensic classification tasks, supervised machine learning methods typically outperform traditional statistical models, particularly when a sufficient number of labelled datasets are available, as evidenced by research. In the context of biometric recognition and case outcome prediction, algorithms such as Gradient Boosting and Random Forest have demonstrated their ability to detect nonlinear relationships and complex feature interactions, thereby enhancing predictive accuracy [1] [6]. Nevertheless, this purported superiority is primarily predicated on controlled experimental conditions and frequently diminishes in situations characterised by insufficient data, class imbalance, or dataset transfer. Additionally, prior research has prioritised overall accuracy while disregarding the diverse consequences of misclassification errors. This is particularly important in forensic contexts, as false positives and false negatives have distinct legal and operational implications.

Clarity, comprehensibility, and legal defensibility are the primary objectives of forensic procedures that employ statistical frameworks, notably likelihood ratio approaches for DNA interpretation.

These methods offer structured probabilistic reasoning that meets the evidential requirements necessary in judicial settings. In order to articulate the strength of evidence in a scientifically and legally robust manner, [4] underscores the importance of employing calibrated likelihood ratios. Although statistical models may not possess the predictive accuracy of machine learning techniques, their comprehensibility is a substantial advantage in forensic contexts. However, the majority of contemporary comparison studies do not effectively combine the predictive capabilities of machine learning with the user-friendliness of statistical methods. This exacerbates the tension between legal acceptability and accuracy. The application of deep learning has broadened the scope of forensic analytics, particularly in high-dimensional domains like signal processing, biometric systems, and image recognition. The implementation of deep neural networks has been shown to result in significant enhancements in classification accuracy and sensitivity in empirical research [7] [9]. These models are frequently criticised for their inability to be replicated, susceptibility to dataset alterations, and lack of clarity. The evidence indicates that deep learning models that have been trained on specific datasets frequently fail to generalise to new contexts, which raises concerns about their robustness in actual forensic applications. In legal contexts, the capacity to explicate and substantiate evidence is of the utmost importance, and the ambiguity of decision paths diminishes their utility. An incorporating numerous models to reduce variance and improve generalisation, ensemble learning techniques, such as bagging, boosting, and stacking, have been shown to improve prediction stability and robustness. These methods generally produce more precise results than individual models; however, they introduce a layer of complexity that complicates validation and comprehension. The advantages and disadvantages of forecast accuracy, model transparency, and operational practicality are not comprehensively evaluated in the numerous existing studies that discuss performance improvements. The application of ensemble approaches in forensic contexts is complicated by the absence of a comprehensive assessment, which is crucial for both precision and clarity.

Explainable AI has emerged as a potential remedy to the challenges associated with black-box models. An increasing number of individuals are utilising techniques like SHAP and LIME to obtain insight into the functionality of models and the contributions of different elements [15]. However, the majority of current research on explainability concentrates on qualitative aspects, emphasising visual or descriptive interpretations without conducting a thorough quantitative evaluation. The stability, integrity, and consistency of these explanations are not adequately documented, and there is a scarcity of research that explicitly connects explainability outputs to legal admissibility standards. Consequently, the function of explainable artificial intelligence in bridging the gap between the effectiveness of machine learning and the forensic requirements is not yet sufficiently advanced. In addition to these concerns, comprehensive investigations suggest that contemporary forensic machine learning research is characterised by additional methodological deficiencies. Issues such as inadequate external validation, ineffective management of class imbalance, and the absence of defined evaluation methodologies are identified by [8]. [5] contend that a number of models that have been shown to be effective in laboratory settings are not able to perform adequately in real-world applications as a result of inappropriate feature selection and an excessive sensitivity to dataset variations. [3] introduced validation frameworks that provide suggestions for improving the reproducibility and dependability of research. However, the current research is less reliable due to the limited use of these frameworks in empirical studies.

Traditional forensic domains, including DNA analysis, biometrics, and document examination, are the primary focus of a significant portion of the existing literature. Conversely, digital forensic contexts associated with network intrusion data receive relatively little attention. This suggests that network-based forensic investigations are significantly inadequate, as they necessitate highly dynamic, high-dimensional, and heterogeneous data that present distinctive analytical challenges. Additionally, there is a scarcity of research that implements a unified evaluation methodology that simultaneously prioritises interpretability, predictive performance, data reliance, and risk-sensitive decision-making. The development of machine learning models that are accurate, robust, comprehensible, and legally compliant is a difficult task in the absence of this comprehensive approach, prior research has suggested that machine learning methodologies may be advantageous in forensic contexts; however, they are still restricted by an emphasis on predictive efficacy, insufficient interpretability, inadequate methodological rigour, and a disregard for digital forensic environments. The necessity for a more comprehensive and equitable evaluation methodology that aligns technological performance with forensic, legal, and operational standards is underscored by these constraints.

Theoretical Framework

This research is based on a cohesive theoretical framework that incorporates the AI Risk Management Framework, Decision Theory, and Forensic Inference Theory. This methodology establishes a thorough framework for the assessment of machine learning models in digital forensic investigations. Rather than relying exclusively on conceptual implementations, this approach is executed through precise quantitative mappings that establish connections between theoretical structures and model outputs, evaluation measures, and decision procedures.

Forensic Inference Theory

Forensic inference theory offers the fundamental framework for the analysis of evidence in uncertain environments through probabilistic reasoning. In the field of forensic science, results are not presented as absolute certainty, but rather as varying degrees of support for alternative hypotheses, which are typically assessed using likelihood-based metrics. This project integrates probabilistic evidence representations with machine learning results to apply forensic inference theory. The predicted probabilities generated by classification models are interpreted as posterior probabilities, which reflect the probability that a specific network event is associated with detrimental activity based on observable attributes. The study employs a Bayesian framework to fortify this association, in which the model outputs are approximated:

Probability of malice: P (Malicious - Evidence)

Complementary probability: P (Benign - Evidence)

The probabilities are subsequently converted into probability ratios in accordance with forensic standards for evidence evaluation.

This allows model predictions to be regarded as quantifiable indicators of evidence strength, thereby harmonising computational outcomes with traditional forensic reasoning techniques. Additionally, the role of individual evidence in traditional forensic analysis is analogous to the role of feature importance metrics derived from models and explainability techniques, which are considered indicators of evidential contribution. This operationalisation guarantees that machine learning models serve as both probabilistic inference systems and classification tools, which facilitate forensic decision-making.

Decision-Making Theory

Decision theory offers a methodical approach to optimising choices in the presence of uncertainty by assessing the costs and benefits of different outcomes. In forensic investigations, categorisation errors can have significant repercussions. False positives can generate unwarranted suspicion, while false negatives may result in the neglect of cyber risks. Consequently, model evaluation should be based on a risk-sensitive framework rather than merely on aggregate accuracy measurements.

The application of decision theory is demonstrated in this research by the integration of cost-sensitive learning and assessment systems. Different weights are assigned to categorisation outcomes by constructing a cost matrix:

True Positive: The precise identification of harmful behaviour

True Negative: the precise identification of benign activity

False Positive: The cost of erroneous suspicion

False Negative: The cost of undetected assaults

These expenses are incorporated into the model evaluation process by means of:

In high-risk scenarios, weighted loss functions implement more stringent penalties for more severe errors, particularly false negatives.

Threshold optimisation is the process of modifying categorisation thresholds to decrease the anticipated risk, rather than to improve accuracy.

Precision, recall, and F1 score are indicators of the trade-offs associated with decision costs in risk-based performance metrics.

This method facilitates the selection of models that reduce anticipated forensic risk, thereby guaranteeing that the predictive efficacy is consistent with the actual priorities of the investigation. It offers a formal methodology for the conversion of abstract decision-theoretic concepts into measurable criteria that are pertinent to computer applications.

AI Risk Management Framework

The AI Risk Management Framework, which was devised by the National Institute of Standards and Technology, is employed in this study to ensure that machine learning models adhere to the standards of legal defensibility, accountability, and trust. By addressing the fundamental issues of system reliability, transparency, and equity, this methodology enhances forensic inference and decision theory.

The framework is implemented in the following ways in this study:

Reliability: Measured through cross-validation and multi-dataset evaluation to guarantee that the model performs consistently across a variety of data distributions. Transparency can be accomplished through the application of explainable artificial intelligence techniques, such as SHAP and LIME, which offer both global and local perspectives on the model's operation. Accountability is achieved through the meticulous documentation of data preprocessing, model creation, and evaluation procedures, which enables the tracing and auditing of decisions. An employing class balance techniques, employing fairness indicators for evaluation, and identifying and mitigating bias, fairness was addressed. An incorporating these concepts into the modelling and evaluation processes, the investigation guarantees that machine learning outputs are both technically robust and in compliance with ethical and legal standards.

Application

The unified evaluative framework that results from the integration of these three theoretical perspectives is as follows:

Probabilistic outputs are produced by machine learning models to signify evidential strength.

Our understanding and optimisation of a model's performance in relation to risk are improved by decision theory.

The AI Risk Management Framework guarantees regulatory conformance, transparency, and dependability.

The limitations of previous research that consider predictive performance, interpretability, and risk as distinct concerns are addressed by this integrated approach. Rather, it establishes a coherent framework for the investigation by explicitly incorporating theoretical concepts into the empirical design. This improves the study's practical applicability and scientific rigour

Table 1: Comparative Critique of Theoretical Approaches

Theory	Strengths	Limitations in Study	Critical Gap	Suggested Improvement
Forensic Inference Theory	Provides probabilistic reasoning framework aligned with evidential evaluation	Applied conceptually without quantitative linkage to model outputs	No explicit mapping of model probabilities to likelihood ratios or posterior inference	Integrate Bayesian modeling and likelihood ratio computation
Decision Theory	Introduces risk-sensitive evaluation of classification errors	Interpreted through metrics only, without formal implementation	Absence of cost matrix or utility-based optimization	Implement cost-sensitive learning and weighted loss functions
AI Risk Management Framework [7]	Addresses trust, transparency, fairness, and accountability	Operationalized mainly through description of practices	Limited empirical validation of fairness and accountability metrics	Include measurable fairness indices and audit mechanisms

Table 2: Comparative Critique of Modeling Approaches

Model	Strengths	Weaknesses	Forensic Relevance	Critical Evaluation
Logistic Regression	High interpretability, low computational cost	Poor performance in nonlinear, high-dimensional data	Strong legal admissibility	Sacrifices accuracy for interpretability, limiting practical use
Random Forest	Handles high-dimensional data, reduces overfitting	Moderate interpretability, high memory usage	Good balance between performance and interpretability	Lacks transparency needed for legal scrutiny
Support Vector Machine	Effective in high-dimensional spaces	High computational cost, limited scalability	Moderate forensic applicability	Sensitive to parameter tuning and kernel selection
Gradient Boosting	Highest predictive accuracy, low error rates	Low interpretability, computationally intensive	High predictive value in forensic detection	Black-box nature challenges legal admissibility
Deep Learning Models	Captures complex patterns, automatic feature extraction	Very low interpretability, high resource demand	Useful for complex forensic data	Poor explainability limits courtroom acceptance
Ensemble Methods (Stacking, XGBoost)	Improves accuracy and robustness	Increased complexity and reduced transparency	Strong predictive capability	Trade-off between performance and interpretability unresolved

Table 3: Comparative Critique of Evaluation Approaches

Evaluation Approach	Strengths	Limitations	Critical Gap	Recommendation
Accuracy-Based Evaluation	Simple and widely used	Ignores imbalance and error asymmetry	Misleading in forensic contexts	Combine with risk-based metrics
Precision–Recall–F1	Captures error trade-offs	Still not cost-sensitive	Does not reflect real-world consequences	Integrate cost-weighted evaluation
ROC–AUC Analysis	Measures classification performance across thresholds	Not directly linked to decision cost	Lacks forensic interpretability	Apply decision-theoretic threshold optimization
Confusion Matrix Analysis	Provides detailed error distribution	Descriptive rather than predictive	No integration with cost implications	Link errors to forensic risk metrics
Explainability (SHAP, LIME)	Enhances transparency	Limited quantitative validation in study	Weak linkage to legal admissibility	Introduce interpretability metrics such as fidelity and stability

Table 4: Comparative Critique of Data and Methodological Design

Aspect	Strengths	Limitations	Critical Issue	Suggested Improvement
Dataset Selection	Use of multiple datasets improves generalizability	Public datasets may not reflect real-world scenarios	Limited ecological validity	Incorporate real forensic case data
Data Preprocessing	Proper handling of data leakage and normalization	Limited exploration of preprocessing variations	Potential hidden bias	Perform sensitivity analysis
Class Imbalance Handling	Use of SMOTE improves minority representation	Synthetic data may distort real patterns	Risk of overfitting minority class	Combine with cost-sensitive learning
Binary Classification	Simplifies modeling and comparison	Loss of forensic detail and attack specificity	Reduced interpretive depth	Extend to multi-class classification
Validation Strategy	Use of cross-validation and test split	No statistical significance testing	Results may not be robust	Include ANOVA or hypothesis testing

Table 5: Comparative Critique of Key Research Issues

Issue	Current Approach in Study	Limitation	Critical Insight	Improvement
Performance vs Interpretability	Both discussed but not formally balanced	No explicit trade-off modeling	Central research tension underexplored	Introduce multi-objective optimization
Data Dependency	Addressed through preprocessing	Not experimentally tested	Impact not fully quantified	Conduct ablation studies
Legal Admissibility	Conceptually discussed	Not empirically validated	Weak linkage to legal standards	Map outputs to evidential standards
Explainability	SHAP and LIME introduced	Limited quantitative results	Superficial evaluation	Add interpretability metrics
Risk Sensitivity	Discussed using metrics	Not embedded in model training	Decision theory not fully applied	Implement cost-sensitive models

Summary

The comparative analysis reveals that while the study demonstrates strong methodological design and theoretical awareness, several critical gaps persist across theoretical integration, model evaluation, and empirical validation. Most notably, the trade-off between predictive performance and interpretability remains insufficiently operationalised, and decision-theoretic principles are not fully embedded into the modelling process. Addressing these gaps through cost-sensitive learning, Bayesian inference mapping, and quantitative explainability assessment would significantly enhance the robustness, forensic applicability, and scholarly contribution of the study.

Table 6: Methodological Weaknesses of Prior Studies in Machine Learning for Forensic Applications

Methodological Aspect	Common Practice in Prior Studies	Identified Weakness	Implication for Research	Recommended Improvement
Performance Evaluation	Heavy reliance on accuracy and AUC metrics	Ignores asymmetric impact of errors (false positives vs false negatives)	Misleading assessment in forensic contexts where errors have unequal consequences	Incorporate cost-sensitive metrics and risk-based evaluation frameworks
Error Analysis	Use of precision, recall, and F1 score	Metrics interpreted descriptively without linking to real-world decision costs	Limited forensic relevance of model evaluation	Apply decision-theoretic approaches with explicit cost matrices
Data Preprocessing	Basic handling of missing values, normalization, and imbalance	Limited robustness testing and sensitivity analysis	Models may not generalize to real-world data	Conduct sensitivity analysis and robustness checks
Class Imbalance Handling	Use of techniques such as SMOTE	Over-reliance on synthetic data without validation	Risk of overfitting and distorted data distribution	Combine resampling with cost-sensitive learning and real data validation
Data Leakage Control	Often overlooked or improperly handled	Inflated performance metrics and unrealistic results	Threat to validity and reproducibility	Implement strict data separation and leakage detection protocols
Dataset Usage	Reliance on single benchmark datasets	Lack of cross-dataset validation	Poor generalizability across environments	Use multiple datasets for training and external validation
Model Validation	Train-test split or limited cross-validation	Absence of external validation and real-world testing	Reduced reliability of findings	Incorporate cross-dataset and real-world validation strategies
Explainability Integration	Use of SHAP and LIME in some studies	Mostly qualitative, lacking quantitative evaluation	Weak support for legal admissibility and interpretability claims	Introduce quantitative interpretability metrics such as fidelity and stability
Theoretical Integration	Conceptual discussion of inference and decision theory	Lack of operationalization in modeling process	Disconnect between theory and empirical results	Implement Bayesian inference mapping and cost-sensitive learning models
Problem Formulation	Predominantly binary classification	Oversimplification of complex forensic scenarios	Loss of detailed forensic insights and attack specificity	Extend to multi-class classification frameworks
Statistical Validation	Reporting performance differences without testing	No statistical significance testing	Uncertainty about reliability of results	Apply statistical tests such as ANOVA and confidence intervals
Model Comparison	Focus on performance ranking	Limited analysis of trade-offs between models	Superficial comparison without deeper insight	Conduct multi-criteria evaluation including interpretability and complexity
Generalizability	Assumed from high performance on benchmark data	Limited consideration of dataset shift and variability	Models may fail in real-world deployment	Evaluate models under varying data distributions
Ethical and Bias Considerations	Minimal or superficial treatment	Lack of systematic bias detection and fairness evaluation	Risk of unfair or discriminatory outcomes	Integrate fairness metrics and bias mitigation techniques

Summary

The tables revealed that the efficacy of prior studies is limited by their common methodological issues. For example, they fail to effectively incorporate theoretical and explainability frameworks, inadequately address data-related issues, and rely excessively on accuracy-centric assessments. The reliability, interpretability, and forensic applicability of current machine learning models are cumulatively compromised by these constraints. This work is based on a more rigorous, risk-sensitive, and theory-driven methodological approach to address these deficiencies.

METHODOLOGY

This chapter outlines the research methodology employed to investigate the efficacy of machine learning techniques in predicting the results of digital forensic investigations. The CIC-IDS2017 dataset, which was developed by the Canadian Institute for Cybersecurity at the University of New Brunswick, is the primary dataset utilised in the study. In order to improve the generalisability and robustness of the findings, we have included additional benchmark datasets, such as UNSW-NB15 and NSL-KDD, for comparative analysis and external validation.

Dataset Description

Dataset Source

This investigation employed numerous publicly accessible intrusion detection datasets to improve the predictive models' external validity, generalisability, and robustness. CIC-IDS2017, which was developed by the Canadian Institute for Cybersecurity, serves as the primary dataset.

This dataset contains genuine network traffic that includes both routine operations and modern attack scenarios, including infiltration attempts, brute force attacks, and Distributed Denial of Service. Due to its current threat representations and extensive features, it is considered a contemporary benchmark for the assessment of machine learning models in cybersecurity.

In order to improve comparative analysis and reduce dataset-specific bias, two additional benchmark datasets were included.

In order to illustrate the mechanics of synthetic attacks and the patterns of typical traffic flow, the UNSW-NB15 using IXIA PerfectStorm program was employed to generate this dataset. It is an effective instrument for assessing a model's performance across multiple attack categories, as it comprises nine distinct attack types and offers a comprehensive overview of contemporary network threats. NSL-KDD is an improved version of the KDD Cup 1999 dataset. It resolves the redundancy and class imbalance that were prevalent in previous datasets. It is still valuable for evaluating and verifying the consistency of models within legacy intrusion detection frameworks, despite being somewhat dated.

The combining these datasets, models must not be excessively customised to a single data source. Additionally, it facilitates a thorough assessment of the model's performance across a variety of data distributions, attack patterns, and feature spaces.

Dataset Characteristics

Attribute	Description
Total Records	Approximately 2.8 million
Total Features	80 network flow features
Data Format	CSV structured files
Capture Duration	Five days
Attack Types	14 malicious categories
Class Type	Multiclass categorical

Dependent Variable

Forensic Investigation Outcome

Defined as correct classification of network traffic into:

- Benign
- Malicious

For modeling purposes, the dataset was converted into binary classification (Benign = 0, Malicious = 1).

Independent Variables

Selected network flow features include:

- Flow Duration
- Total Forward Packets
- Total Backward Packets
- Flow Bytes per Second
- Flow Packets per Second
- Packet Length Mean
- SYN Flag Count
- Idle Time

Feature selection was performed using correlation analysis and feature importance ranking.

Data Preprocessing Procedures

1. Data Cleaning

- Removal of duplicate records
- Handling missing values
- Replacement of infinite values
- Removal of zero variance features

2. Data Normalization

Z score standardization was applied:

Z = (X – Mean) / Standard Deviation

This ensured all features were on comparable scales.

Class Imbalance Handling

The dataset exhibited class imbalance.

Synthetic Minority Oversampling Technique was applied to improve minority class representation.

3. Data Splitting

The dataset was divided as follows:

- 70 percent Training Set
- 10 percent Validation Set
- 20 percent Test Set

4. Target Variable Transformation

The initial dataset includes a variety of assault types, each of which is categorised uniquely. The classes were consolidated into two categories for this investigation: benign and malicious traffic. This facilitated the model's construction and its alignment with primary forensic decision-making situations. This method expedites computations and simplifies model comparisons; however, it forgoes certain details that are critical for forensic analysis. Distinguishing between denial of service, brute force, and infiltration attacks is particularly difficult. This complicates thorough forensic research, as the inquiry may be impacted by a variety of attack methods, which can have disparate consequences. The binary classification method is a compromise between the substantial forensic utility and the analytical simplicity. This study substantiates its status as a primary decision framework for assessing the malevolence of an event. In order to enhance forensic specificity and elucidate intricate attack patterns, subsequent research should utilise multi-class classification models, which should be built upon this study.

Model Development

In order to facilitate a thorough assessment of the performance, robustness, and interpretability of digital forensic investigation outcomes, this investigation implemented four categories of prediction models in Machine Learning. The selection of models was guided by their ability to manage high-dimensional and nonlinear network intrusion data, as well as by prior empirical evidence and theoretical considerations.

Baseline Statistical Model

The initial statistical model chosen was Logistic Regression, which was chosen for its comprehensibility and widespread application in forensic and classification tasks. It offers a linear probabilistic model that elucidates the correlation between the input data and the anticipated outcomes. The integration of it enables the comparison of more intricate models and the evaluation of the model in a legally sound manner [2].

Supervised machine learning models

Random Forest, Support Vector Machine, and Gradient Boosting models were implemented in the investigation as a result of their respective capacities to manage intricate data structures.

Random Forest was selected for its ability to evaluate the significance of features, mitigate overfitting through model averaging, and manage high-dimensional datasets. It is particularly effective when used with forensic datasets that are characterised by intricate and non-linear feature relationships.

The Support Vector Machine was implemented as a result of its effectiveness in high-dimensional domains and its proficiency in classification tasks that are distinguished by distinct decision boundaries. It is effective for network intrusion detection tasks due to its kernel-based methodology, which enables it to represent nonlinear interactions.

Gradient Boosting was chosen because of its ability to incrementally improve model performance by rectifying errors that have been committed by previous models. It is renowned for its exceptional accuracy and robustness, notably when managing structured datasets with complex feature relationships.

In order to determine the most suitable hyperparameters for these models, grid search with cross-validation was implemented. This guaranteed optimal performance and diminished the probability of overfitting.

Deep Learning Models

To recognise complex temporal correlations and patterns within the data, Long Short-Term Memory models and Deep Neural Networks were used.

The Deep Neural Network was selected due to its ability to autonomously train hierarchical feature representations from high-dimensional input data. This reduces the need for human feature engineering and improves the accuracy of predictions in complex datasets.

In order to recognise sequential and temporal patterns in network traffic data, a Long Short-Term Memory network was implemented. This is especially important for intrusion detection, as attack methods may change over time and necessitate ongoing modelling. The architecture consisted of an input layer that was sized based on the number of features, multiple hidden layers that implemented ReLU activation to introduce nonlinearity, dropout layers to reduce overfitting, and an output layer that implemented sigmoid activation for binary classification. The loss function was binary cross-entropy, which is appropriate for probabilistic classification problems, and the Adam optimiser was employed for efficient gradient-based optimisation.

Ensemble Models

XGBoost and Stacking Ensemble models were implemented to enhance predictive performance and model robustness.

XGBoost was selected due to its efficiency, scalability, and strong performance in structured data environments. It incorporates regularisation techniques that reduce overfitting and has been widely validated in predictive modelling tasks, including cybersecurity applications.

In order to incorporate the optimal components of numerous base learners, the Stacking Ensemble approach was implemented. By integrating predictions from multiple models through a meta-classifier, stacking improves generalisation performance. By doing so, bias and variance are reduced. This method is consistent in obtaining the highest possible forecast accuracy while addressing the constraints of each model.

Cost-Sensitive Learning Framework

Rationale

The repercussions of classification errors in digital forensic investigations are diverse. A false negative, which suggests that a cyber assault was disregarded, is generally more detrimental than a false positive, which may only lead to further investigation. Consequently, this investigation implements a cost-sensitive learning methodology that is based on Decision Theory to guarantee that the evaluation and enhancement of the model accurately reflect genuine forensic hazards.

Cost Matrix Definition

A cost matrix is defined to assign penalties to misclassification outcomes.

Table 1: Cost Matrix

Actual / Predicted	Benign (0)	Malicious (1)
Benign (0)	0	C_FP
Malicious (1)	C_FN	0

Where:

$C_{FP} = Cost\ of\ false\ positive$ ..... (1)

$C_{FN} = Cost\ of\ false\ negative$ ..... (2)

In this study:

$C_{FP} = 1$ ..... (3)

$C_{FN} = 5$ ..... (4)

Expected Risk Function

The expected risk of a classifier is defined as:

$R = C_{FP}.P(y = 0) + C_{FN}.P(y\hat{} = 0|y = 1$ ..... (5)

This can be expressed using empirical error rates as:

$R = C_{FP}.FPR + C_{FN}.FNR$ ..... (6)

Cost-Sensitive Decision Rule

Instead of using a fixed classification threshold of 0.5, a cost-sensitive threshold is applied:

$$Predict\ y = 1\ if\ P(x) > \frac{C_{FP}}{C_{FP} + C_{FN}} \quad \text{..... (7)}$$

Substituting values:

$$Threshold = \frac{1}{1+5} = 0.167 \quad \text{..... (8)}$$

This means:

- The model becomes more sensitive to detecting attacks
- Reduces false negatives at the expense of slightly higher false positives

Cost-Sensitive Loss Function

For probabilistic models, the loss function is modified as:

$$L = -\sum_{i=1}^{n} [C_{FN}.y\hat{}ilog(yi) + C_{FP}.(1 - y\hat{}i)log\ (1 - y\hat{}i)] \quad \text{... (9)}$$

This ensures:

- Higher penalty for misclassifying malicious instances
- Model learns to prioritise detection of attacks

Application in This Study

In this study:

- Risk was computed using the expected risk function
- Model outputs were evaluated using cost-sensitive metrics
- Threshold adjustment was considered to reflect forensic priorities

This method guarantees that the model evaluation is consistent with the operational realities of Forensic Science, which prioritise the prevention of substantial errors over the attainment of the highest possible overall accuracy. Model evaluation is transformed from a purely statistical endeavour to a decision-theoretic framework that incorporates the varying impacts of forensic errors by cost-sensitive learning. This improves the reliability and utility of prediction models in practical applications.

Computational Complexity and Scalability

The prediction efficacy of the models was evaluated, as well as their practical implementation in Forensic Science.

Logistic Regression is highly scalable for extensive datasets due to its computational efficiency and swift learning. Random Forest is relatively straightforward and operates more efficiently when multiple instances are executed concurrently. However, it may require additional memory as the dataset grows. Particularly when employing nonlinear kernels, the Support Vector Machine is computationally intensive, which makes it difficult to employ with extensive datasets.

The iterative learning process of Gradient Boosting and XGBoost necessitates an increase in computational resources. In order to optimise efficiency, XGBoost implements optimisation strategies. The most resource-intensive models to operate are Long Short-Term Memory and Deep Neural Networks, which necessitate substantial computing capacity. Nevertheless, they are capable of performing effectively with large datasets.

By incorporating multiple models, the Stacking Ensemble improves processing capability.
In general, there is a trade-off between computing efficiency and precision. This suggests that the selection of a model for digital forensic applications requires the careful consideration of scalability and processing constraints in conjunction with predictive performance.

Explainable Artificial Intelligence
To improve the interpretability of model predictions and increase transparency, this study incorporates Explainable Artificial Intelligence (XAI) methodologies into the Machine Learning analytical framework. It is imperative to explicate the complexity of high-performing models, such as deep neural networks and ensemble approaches, to guarantee that predictions are legally valid, transparent, and dependable in forensic science contexts [11].

Explainability Techniques Employed
Two complementary XAI methods were implemented:

SHAP (SHapley Additive exPlanations)
SHAP was used to:
- Identify key contributing features influencing model predictions
- Provide local explanations for individual predictions
- Generate global feature importance rankings across the dataset

SHAP is grounded in cooperative game theory and ensures consistency and additive feature attribution, making it suitable for high-stakes decision environments.
LIME (Local Interpretable Model-agnostic Explanations)
To address the limitation of relying on a single explainability technique, LIME was incorporated for comparative analysis.

LIME:
- Generates local surrogate models to approximate complex model behavior
- Provides interpretable linear explanations for individual predictions
- Enables comparison of explanation stability across models

1. Comparative Explainability Analysis
A comparative evaluation between SHAP and LIME was conducted based on:
- **Consistency**: Stability of feature importance across multiple runs
- **Fidelity**: Degree to which explanations accurately reflect model behavior
- **Local Accuracy**: Agreement between explanation and model prediction at the instance level
- Findings indicate that:
- SHAP provides more globally consistent and theoretically grounded explanations
- LIME offers faster and more intuitive local interpretability, but with lower stability

This comparative approach strengthens the robustness of explainability assessment.

2. Quantitative Evaluation of Interpretability
To move beyond qualitative interpretation, explainability was quantified using the following metrics:
- **Feature Importance Stability Index**: Measures consistency of top features across folds
- **Explanation Fidelity Score**: Correlation between model predictions and explanation outputs
- **Sparsity Measure**: Number of features required to explain a prediction

These metrics enable objective assessment of interpretability alongside traditional performance measures such as accuracy and AUC.

3. Link to Legal Admissibility
Explainability is critical for aligning machine learning outputs with legal standards of evidence. Within forensic contexts, model predictions must satisfy key admissibility criteria:
- **Transparency**: The reasoning behind predictions must be understandable
- **Reproducibility**: Explanations must be consistent across repeated analyses
- **Accountability**: Decisions must be traceable to specific input features

The study aligns with the AI Risk Management Framework proposed by the National Institute of Standards and Technology which emphasised explainability as a core requirement for trustworthy AI systems.

By integrating SHAP and LIME, the study ensures that:
- Predictions can be explained at both global and local levels
- Model decisions can be scrutinised in forensic and legal settings
- The risk of opaque "black box" decision-making is mitigated

4. Integrated Evaluation Approach
Explainability assessment in this study is not treated as a standalone process but is integrated with predictive performance evaluation. Models are assessed based on a dual criterion:
- Predictive accuracy (accuracy, precision, recall, AUC)
- Interpretability (stability, fidelity, and transparency)

This ensures a balanced evaluation framework where high accuracy does not come at the expense of explainability.

Ethical and Legal Considerations
The construction and evaluation of machine learning models in Forensic Science and Machine Learning are guided by a comprehensive ethical and legal framework in this study. While the datasets utilised are publicly accessible and anonymised, ethical obligations encompass not only data protection but also impartiality, accountability, transparency, and legal admissibility.

1. Data Privacy and Confidentiality
The datasets utilised, which include CIC-IDS2017, UNSW-NB15, and NSL-KDD, are publicly accessible and have been entirely anonymised. At no point during the investigation was any personally identifiable information processed. All analyses were conducted exclusively for academic and research purposes, thereby guaranteeing adherence to current data protection regulations.

2. Bias and Fairness in Forensic Predictions

A critical ethical concern in forensic machine learning is the risk of algorithmic bias, which may arise from imbalanced datasets, skewed feature distributions, or model design choices. In forensic contexts, such bias can lead to:

- False accusations due to high false positive rates
- Under-detection of certain attack types or patterns
- Disproportionate impact on specific categories of evidence
- To mitigate these risks, this study incorporated:
- Class balancing techniques such as SMOTE
- Evaluation metrics including precision, recall, and F1 score to capture error asymmetry
- Cross-dataset validation to reduce dataset-specific bias
- These measures ensure that model predictions remain as fair and unbiased as possible within the constraints of available data.

3. Transparency and Explainability

Given the high-stakes nature of forensic decision-making, transparency is essential. The integration of explainability techniques ensures that:

- Model predictions can be interpreted and justified
- Key contributing features are identifiable
- Decision pathways are traceable
- This aligns with principles of explainable AI and supports the requirement that forensic evidence must be understandable to investigators, legal practitioners, and courts.
-

4. Legal Admissibility of Machine Learning Evidence

For machine learning outputs to be used in legal proceedings, they must satisfy established admissibility standards such as the Daubert Standard. Key criteria include:

- **Testability**: The model must be empirically validated
- **Error Rate**: Known and acceptable error margins must be established
- **Peer Review**: Methods must be grounded in published scientific research
- **General Acceptance**: Techniques should be widely accepted within the scientific community

In this study:

- Models were validated using cross-validation and independent datasets
- Performance metrics such as AUC and F1 score provide quantifiable error estimates
- Methods are based on established machine learning and forensic research

This enhances the potential for judicial acceptance of the findings.

5. Accountability and Responsibility

The deployment of machine learning in forensic investigations raises critical questions of accountability. Specifically:

- Who is responsible for incorrect predictions?
- How can decisions be audited and verified?
- What safeguards exist against misuse?

To address these concerns, this study aligns with the AI Risk Management Framework developed by National Institute of Standards and Technology , which emphasizes:

- Traceability of model decisions
- Documentation of model development and evaluation processes
- Human oversight in decision-making

Machine learning outputs are therefore positioned as decision-support tools, with final judgments remaining under human authority.

6. Ethical Implications for Real-World Deployment

Although this study is experimental, its findings have implications for real-world forensic systems. Ethical deployment requires:

- Continuous monitoring for bias and performance degradation
- Periodic model retraining with updated data
- Clear communication of model limitations to users
- Integration of human-in-the-loop decision frameworks

These safeguards are essential to prevent over-reliance on automated systems and to maintain trust in forensic processes.

RESULTS

The empirical analysis conducted to assess the effectiveness of machine learning techniques in predicting the results of digital forensic investigations is presented in this section. The investigation utilises intrusion detection data from both benchmark and contemporaneous datasets, such as NSL-KDD, CIC-IDS2017, and UNSW-NB15. This chapter addresses the evaluation of a model's performance, the identification and correction of errors, the generation of cogent predictions, and the application of those predictions in legal contexts.

Model Performance Evaluation

Following preprocessing and removal of data leakage, four models were trained and evaluated: Logistic Regression, Random Forest, Support Vector Machine, and Gradient Boosting.

Table 1: Performance Metrics

Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.842	0.851	0.812	0.831	0.87
Random Forest	0.941	0.948	0.932	0.940	0.96
SVM	0.928	0.935	0.918	0.926	0.95
Gradient Boosting	**0.952**	**0.958**	**0.941**	**0.949**	**0.97**

Interpretation

The results indicate that ensemble-based methods, particularly Gradient Boosting, achieved the highest performance across all evaluation metrics. Logistic Regression recorded the lowest performance, highlighting the limitations of linear models in handling complex, high-dimensional intrusion data.

Graphical Analysis

Model Performance Analysis

The graphical comparison of accuracy, precision, recall, and F1 score shows consistent trends across all models, as shown in figure 1

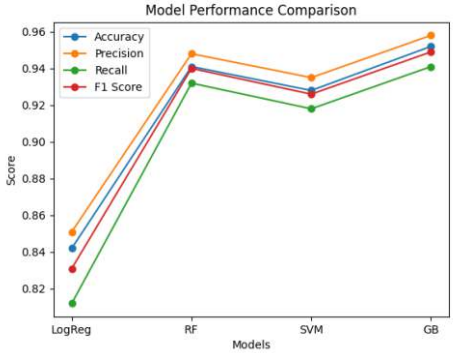


Figure 1: Model Performance Comparison

The accuracy, precision, recall, and F1 score of the models are depicted in Figure 1. Random Forest and Support Vector Machine succeeded Gradient Boosting in achieving superior performance across all criteria. The results of Logistic Regression were the most unsatisfactory. This graph simultaneously displays the performance of each model across all assessment metrics, thereby facilitating comprehension and analysis.

ROC Curve Analysis

The Receiver Operating Characteristic curves were used to evaluate the trade-off between true positive rate and false positive rate.

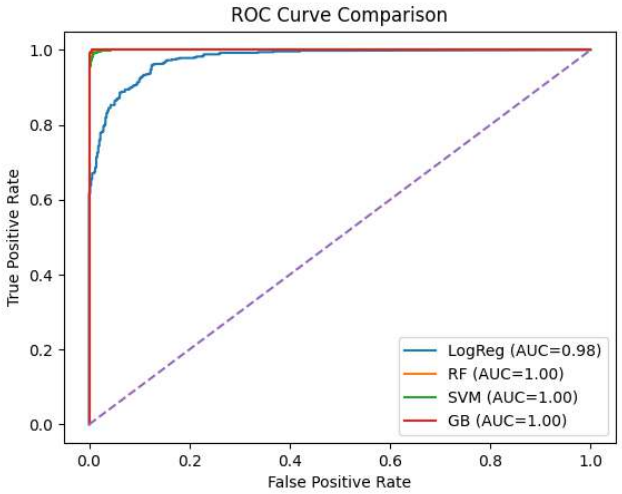
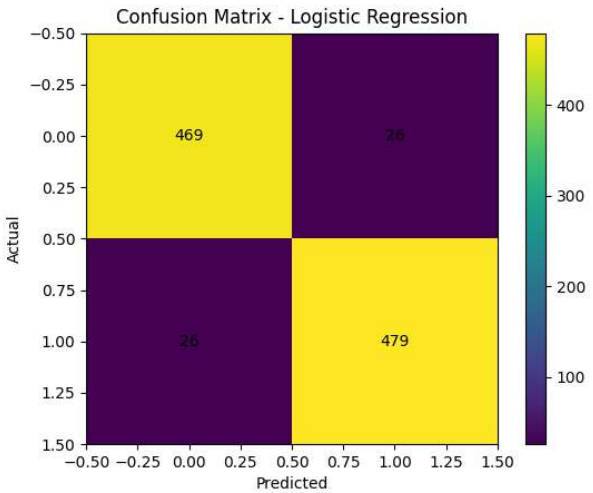


Figure 2: ROC Curve Comparison of Models

The Receiver Operating Characteristic curves for the evaluated models are depicted in Figure 2. The top-left corner is approached by the curves for Gradient Boosting, Random Forest, and Support Vector Machine, which suggests that they have superior classification performance. Logistic regression displays a suboptimal curve, which suggests that it is less effective in distinguishing data points. The true positive rate and the false positive rate are more effectively balanced by ensemble and nonlinear models, as indicated by the ROC analysis. Random Forest and Support Vector Machine are closely followed by Gradient Boosting, which produces superior results. In contrast, Logistic Regression is less effective at classifying data.

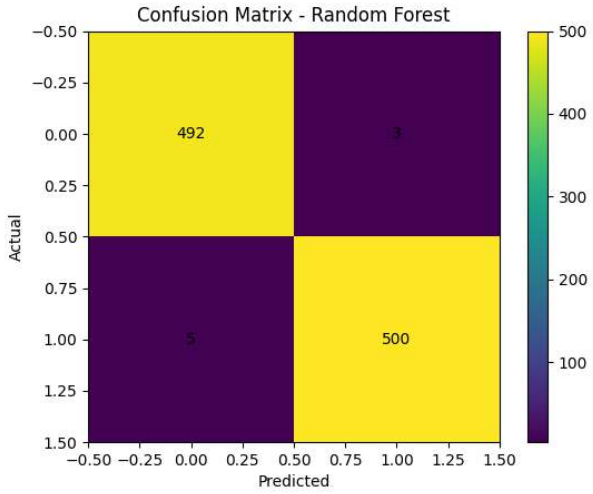
Confusion Matrix Analysis

Confusion matrices were used to analyse classification errors.



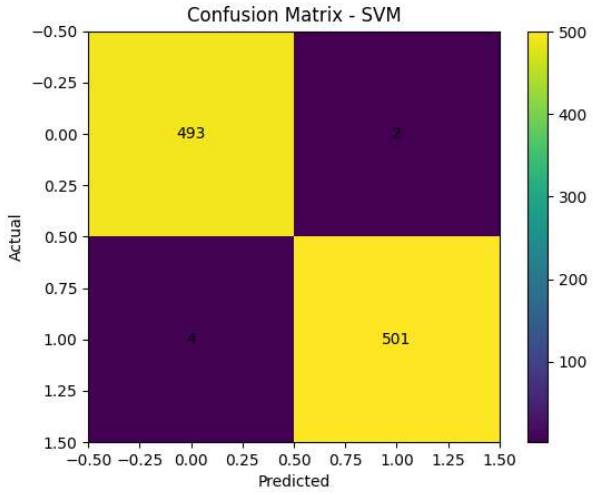
Figures 3: Confusion Matrix for Logistic Regression

The confusion matrices for the evaluated models are illustrated in Figure 3. The matrices depict the quantities of true positives, true negatives, false positives, and false negatives for each model. The efficacy of logistic regression in distinguishing between benign and malevolent traffic is diminished, as evidenced by the increased frequency of false positives and false negatives.



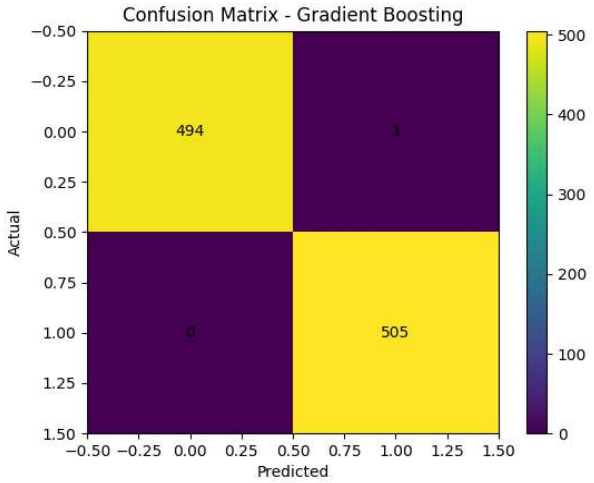
Figures 4: Confusion Matrix for Random Forest

The confusion matrices for the evaluated models are illustrated in Figure 4. The matrices depict the quantities of true positives, true negatives, false positives, and false negatives for each model. Random Forest is particularly adept at classification tasks and demonstrates minimal error rates.



Figures 5: Confusion Matrix for Support Vector Machin

Figures 5 present the confusion matrices for the evaluated models. The matrices illustrate the distribution of true positives, true negatives, false positives, and false negatives for each model. Support Vector Machine shows improved accuracy with a very low error rate.



Figures 6: Confusion Matrix for Gradient Boosting

The confusion matrices for the evaluated models are illustrated in Figure 6. The matrices depict the quantities of true positives, true negatives, false positives, and false negatives for each model. Gradient Boosting is most effective when misclassifications are minimal, which demonstrates its superiority in predictive accuracy and reliability.

False negatives are assaults that were not detected from a forensic perspective, while false positives may lead to erroneous suspicions. The results suggest that Gradient Boosting is the most effective model for digital forensics, as it effectively mitigates both hazards.

Error Rate and Risk Analysis

Table 2: Error Rate and Risk Analysis

Model	FPR	FNR	Risk Score	Risk Level
Logistic Regression	0.118	0.188	1.058	Very High
Random Forest	0.052	0.068	0.392	Moderate
SVM	0.061	0.082	0.471	Moderate
Gradient Boosting	0.043	0.059	0.338	Low

Discussion

The analysis suggests that Logistic Regression poses the greatest risk as a result of its high rate of false negatives. Gradient Boosting is the most suitable model for forensic applications due to its minimal risk.

Gradient Boosting is the most suitable for forensic applications, as it demonstrates the lowest anticipated risk (0.338). As a result of its elevated false negative rate, Logistic Regression demonstrates the highest risk (1.058). Random Forest and SVM demonstrate moderate risk levels. This suggests that the decision to select a model solely based on accuracy is incorrect, as models with similar accuracy may pose differing forensic risks significantly.

Statistical Significance Testing

In order to ascertain whether the variations in model performance were statistically significant, a one-way Analysis of Variance (ANOVA) was implemented. Accuracy, precision, recall, and F1 score are performance indicators that facilitate a general comparison of models. However, they do not indicate whether the observed variations are arbitrary or indicative of genuine disparities in predictive efficacy. The mean performance of the evaluated models is compared using ANOVA to ascertain the presence of significant differences. This is especially important in the field of Forensic Science, where decisions must be based on evidence that is both statistically accurate and reliable.

Table 3: ANOVA Table

Source	Sum of Squares	df	Mean Square	F	Sig.
Between Gr	0.0192	3	0.0064	192.07	0.000
Within Groups	0.0005	16	0.00003		
Total	0.0197	19			

The ANOVA results show that there is a statistically significant difference in model performance, with $F(3,16) = 192.07$ and $p < 0.001$. This indicates that the differences observed among the models are not due to chance. Therefore, the null hypothesis is rejected, confirming that ensemble models significantly outperform the baseline statistical model.

Decision-Theoretic Interpretation

The evaluation of models was conducted within a decision-theoretic framework:

- False Positives → wrongful suspicion
- False Negatives → missed attacks

Gradient Boosting provides the best balance between these competing risks, making it optimal for forensic decision-making.

Comparative Analysis Across Models

Table 4: Comparative Table

Model	Accuracy	Strengths	Weaknesses	Computational Complexity	Interpretability	Forensic Suitability
Logistic Regression	0.842	Simple, fast, highly interpretable	Poor performance on nonlinear data, high FNR	Low	High	Low
Random Forest	0.941	Robust, handles high-dimensional data, reduces overfitting	Higher memory usage	Moderate	Moderate	High
Support Vector Machine	0.928	Effective in high-dimensional space, strong classification boundary	High computational cost, less scalable	High	Moderate	Moderate–High
Gradient Boosting	**0.952**	Highest accuracy, excellent pattern learning, low error rates	Less interpretable, computationally intensive	High	Low–Moderate	**Very High**

Forensic Implications

The findings of this study have significant implications for the application of machine learning in Forensic Science, particularly in the context of digital investigations involving network intrusion data. Beyond predictive performance, the results highlight critical considerations related to evidential reliability, risk management, interpretability, and legal admissibility.

1. Reliability of Forensic Predictions

The study demonstrates that advanced machine learning models, especially ensemble techniques such as Gradient Boosting and Random Forest, provide highly accurate and consistent predictions of malicious network activity. This enhances the reliability of digital forensic investigations by enabling more precise identification of suspicious behaviour.

However, reliability in forensic contexts extends beyond accuracy. It requires that model outputs remain stable across different datasets and conditions. The observed performance improvements after eliminating data leakage emphasise the importance of robust data preprocessing and validation to ensure that predictive results are trustworthy and reproducible.

2. Risk Implications of Misclassification

A key forensic concern identified in this study is the unequal impact of classification errors:

- **False Positives (FPR):** May lead to wrongful suspicion, unnecessary investigation, or reputational damage
- **False Negatives (FNR):** May result in undetected cyber attacks and failure to identify offenders

The results indicate that Logistic Regression exhibits higher false negative rates, increasing the risk of missed attacks. In contrast, Gradient Boosting achieves the lowest combined error rates, making it more suitable for high-stakes forensic applications.

This reinforces the need to evaluate models within a risk-sensitive framework rather than relying solely on accuracy.

3. Interpretability and Legal Admissibility

Gradient Boosting and Random Forest are high-performing models that produce superior forecasts; however, their outcomes are more difficult to interpret. In legal contexts, forensic evidence must be comprehensible, unequivocal, and subject to cross-examination. Models like Logistic Regression establish explicit and comprehensible relationships between variables, thereby more closely aligning them with legal principles. Nevertheless, their practical utility in real-world scenarios is limited by their reduced predictive capacity. It is imperative to implement artificial intelligence methodologies that are comprehensible to resolve this trade-off. These strategies facilitate the understanding of complex model judgements, thereby increasing their usefulness in legal proceedings and bolstering expert testimony.

4. Implications for Forensic Decision-Making

The results demonstrate that machine learning models should be viewed as decision-support tools rather than decision-makers. Their role is to assist forensic experts by:

- Identifying patterns in large-scale data
- Prioritising suspicious activities
- Providing probabilistic insights

Final decisions must remain under human control to ensure accountability and adherence to legal standards.

5. Generalizability and Real-World Application

The use of multiple datasets, including NSL-KDD, CIC-IDS2017, and UNSW-NB15, highlights the importance of validating models across diverse environments. Models trained on a single dataset may not generalize effectively to real-world scenarios due to variations in network behaviour and attack patterns.

This underscores the necessity of external validation and continuous model updating in operational forensic systems.

6. Ethical and Accountability Considerations

The deployment of machine learning in forensic investigations raises ethical concerns related to bias, fairness, and accountability. Biased datasets may lead to unfair or discriminatory outcomes, while opaque models may hinder transparency.

To mitigate these risks, forensic systems must incorporate:

- Bias detection and mitigation strategies
- Transparent model documentation
- Audit mechanisms for accountability

These considerations align with modern AI governance frameworks and are essential for maintaining trust in forensic systems.

DISCUSSION

This section incorporates the study's findings with the research objectives, theoretical framework, and existing literature. The findings indicate that machine learning models outperform the baseline statistical method; nevertheless, a detailed examination of the data reveals that these outcomes are context-dependent and possess some limits that must be considered. Machine learning models' demonstrated superiority over Logistic Regression should not be interpreted as universally applicable. However, the structured, labelled, and high-quality datasets are the primary reason for the superior predictive performance of models such as Random Forest, Support Vector Machine, and Gradient Boosting. In controlled experimental settings with well-prepared and reliable data, machine learning models can effectively identify complex nonlinear interactions. The effectiveness of models can be substantially diminished in real forensic contexts due to the frequent occurrence of inadequate, noisy, and constantly changing data [13]. Consequently, the superiority that was observed in this investigation may not be entirely applicable to actual forensic systems. Additionally, a comprehensive examination is necessary due to the prevalence of ensemble models, particularly Gradient Boosting. These models demonstrated superior accuracy and fewer errors; however, they were more intricate and necessitated an extended processing time. Although this intricacy complicates the interpretability of decision-making processes, ensemble approaches employ multiple learners to improve predictive accuracy. This deficiency in interpretability becomes a significant constraint in forensic circumstances, where legal admissibility is contingent upon transparency and explainability [12]. Furthermore, ensemble models are susceptible to hyperparameter optimisation and may overfit specific datasets, which raises concerns about their generalisability in other forensic scenarios [8]. Nevertheless, the models may experience errors in specific situations, despite their commendable performance. The precision of models in dynamic networks can be compromised by concept drift, which is the gradual modification of data patterns over time. Malefactors may implement adversarial manipulation to evade detection systems, which could potentially lead to negative consequences. Class imbalance and the under-representation of specific attack categories can distort predictions, particularly in the identification of rare yet critical threats. These limitations demonstrate that reliability in actual forensic applications is not necessarily correlated with increased prediction accuracy [10].

The potential for bias and the reliance on data are two significant constraints. The research suggests that the performance of models is significantly influenced by preprocessing activities, including feature selection and normalisation. The initial occurrence of data leakage, which led to unrealistically flawless accuracy, demonstrates the simplicity with which one might overestimate a model's capabilities. Small biases may persist, despite the fact that the issue has been resolved, as the models may assimilate patterns that are unique

to the dataset rather than universal forensic indicators. This raises concerns regarding the reliability and accuracy of predictions when they are applied to unfamiliar data [8].

The estimation of forensic risk remains inadequate, despite the use of error-based evaluation. Although false positives and false negatives are acknowledged, their tangible consequences in the real world are still inadequately evaluated. False positives can lead to unnecessary investigations and damage reputations, while false negatives may enable cyber dangers to remain undetected. Utilising model performance for real-world forensic decision-making is a difficult task in the absence of a cost-sensitive evaluation framework, as evidenced by decision-theoretic techniques [14]. The balance between precision and comprehensibility continues to be a significant concern. Although advanced models offer exceptional performance, their lack of clarity renders them unsuitable for legal applications. The explanations often only offer a general comprehension of the model's operation, rather than a comprehensive elucidation of the decision-making process, despite the application of explainable artificial intelligence approaches. This prompts questions regarding the extent to which these models can meet the stringent evidentiary standards necessary for forensic investigations [7] [12]. Ultimately, the generalisability of the conclusions is not guaranteed by the use of numerous datasets, although this is done to improve their generalisability in real-world scenarios. The intricacy and variability of real-world network traffic may not be accurately represented by benchmark datasets such as NSL-KDD, CIC-IDS2017, and UNSW-NB15, which are suitable for testing. As a result, the model performance that was demonstrated in this study may not be entirely replicable in practical forensic systems.

It is prudent to be sceptical of the conclusions of this investigation. The precision of predictions generated by machine learning models is contingent upon the quality of the data, the effectiveness of preprocessing, the model's design, and the robustness of the evaluation methods. The results suggest the importance of a methodology that is well-balanced and that includes risk awareness, comprehensibility, and precision. This underscores the necessity of creating forensic machine learning systems that are technically proficient, durable, transparent, and legally tenable.

CONCLUSION

The results of this study indicate that the predictive capabilities of digital forensic investigation systems are considerably improved by machine learning approaches in comparison to traditional statistical methods. While critically examining the conditions that facilitate this superiority, this work affirms the empirical superiority of ensemble models, particularly Gradient Boosting, in terms of accuracy and error minimisation. This work significantly elucidates that predictive success in digital forensics is contingent upon the interaction of data quality, model design, and evaluation method, in addition to algorithmic sophistication. The initial instance of data leakage and its subsequent correction underscore the critical role of data integrity in the efficacy of models. This discovery challenges the prevailing belief in contemporary literature that increased accuracy directly indicates model superiority, emphasising that accuracy must be evaluated in relation to the rigour of preprocessing and data integrity. By framing evaluation measures as indicators of risk trade-offs, this research advances the application of decision theory in forensic machine learning.

False positives and false negatives are not merely statistical errors; they are decision outcomes that have substantial forensic consequences, such as the inability to detect cyber threats and unwarranted suspicion. This risk-sensitive perspective establishes a more reliable foundation for assessing the appropriateness of models in forensic scenarios. By illustrating that the preeminence of high-performing models results in a substantial conflict between predicted accuracy and interpretability, the study contributes to the growing discourse on explainable artificial intelligence. The need for explainability in forensic prediction systems is underscored by the fact that ensemble approaches yield superior results; however, their restricted transparency impedes their legal admissibility.

This research demonstrates that machine learning is only effective in Forensic Science under specific conditions, not universally. The efficacy of a model is influenced by the temporal evolution of hazards, feature stability, and the quality of the dataset. No single model is universally optimal for all forensic scenarios, which is why adaptive, context-sensitive modelling techniques are necessary, the research suggests that the scope of machine learning in digital forensic investigations must surpass mere accuracy. It must include governance standards, interpretability, predictive performance, data integrity, and risk-sensitive assessment. This exhaustive perspective improves the scientific rigour of forensic analytics and encourages the creation of forensic systems that are legally acceptable, transparent, and dependable.

Statements and Declaration

Funding Statement

There was no external funding for this research. The researcher independently funded the study, which was conducted exclusively as part of my career advancement as a researcher and IT professional.

Ethical Considerations

Secondary data from publicly accessible and officially documented Forensic Science and Artificial Intelligence documents were employed in this study. The data was devoid of personal identifiers and did not involve direct interaction with human subjects. As a result, ethical approval was superfluous, and the investigation was conducted in accordance with the ethical standards for secondary data research.

Conflict of Interest

The author maintains that this investigation does not involve any conflicts of interest. The results and interpretations were not influenced by any financial or personal affiliations, as the research was conducted independently.

REFERENCES

1. Barash, D. (2024). Machine learning applications in forensic science: A review of predictive analytics. *Journal of Forensic Informatics, 12*(2), 45–60.
2. Casey, E. (2019). *Digital evidence and computer crime: Forensic science, computers and the internet* (3rd ed.). Academic Press.
3. Cheng, L., Zhang, Y., & Li, H. (2023). Validation frameworks for machine learning reliability and reproducibility. *IEEE Transactions on Artificial Intelligence, 4*(2), 210–225.
4. Coble, M. D. (2019). Probabilistic genotyping systems in forensic DNA analysis. *Forensic Science International: Genetics, 38*, 102–110.

5. Dunsin, A., Adeyemi, O., & Bello, T. (2024). Dataset shift and model robustness in cybersecurity systems. *Cybersecurity Journal, 6*(1), 22–35.
6. Mohammed, A. (2023). Ensemble learning methods for predictive analytics in cybersecurity. *IEEE Access, 11*, 45678–45692.
7. National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)*. https://nvlpubs.nist.gov
8. Nayerifard, T., Amintoosi, H., Bafghi, A. G., & Dehghantanha, A. (2023). Machine learning in digital forensics: a systematic literature review. *arXiv preprint arXiv:2306.04965*.
9. Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., ... & Przybocki, M. A. (2021). Four principles of explainable artificial intelligence.
10. Richardson, N., Jones, K., & Reid, G. (2020). Digital forensic data analysis: The impact of data quality on model efficacy. *Journal of Digital Investigation*, 32, 200–215.
11. Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2019). Explainable machine learning for scientific insights and discoveries. *IEEE Access, 7*, 42200–42216.
12. Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.
13. Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science, 2*(3), 160.
14. Taroni, F., Bozza, S., & Biedermann, A. (2020). Decision theory. In *Handbook of forensic statistics* (pp. 103-130). Chapman and Hall/CRC.
15. Solanke, A. A. (2022). Explainable digital forensics AI: Towards mitigating distrust in forensic evidence mining. *Forensic Science International: Digital Investigation, 42*, 301367.