National Research Journal of Information Technology & Information Science
Volume No: 12, Issue No: 2, Year: 2025 (July- December)
PP: 360-376
ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

AN INTERPRETABLE AND OPTIMIZED HYBRID MACHINE LEARNING FRAMEWORK FOR DATA PRIVACY THREAT DETECTION AND ETHICAL MODEL EVALUATION

Parveen Kumar Goyal

Research Scholar, School of Computer Application, Career Point University, Kota, Rajasthan, India

Garima Tyagi

Professor, School of Computer Application, Career Point University, Kota, Rajasthan, India

ABSTRACT

This work investigates the urgent need for models that are simultaneously robust and responsible in Data Privacy Threat detection. In this paper a Hybrid Machine Learning Framework is to be used that fuses Convolutional Neural Network for feature learning, also an XGBoost classifier to be implemented which has been carefully optimized using optuna bayesian approach. It resulted in better classification performance with an AUC 0.999 and F1-Score of 0.9886, clearly outperforming unoptimized baselines. Importantly, conceptualization is taken beyond mere performance, such as by introducing a way to measure model interpretability and ethical reasoning. Hear a new comparative study quantifying important operational statistics such as and the Traceability Index to steer deploy resource efficiency, leverage Natural Language Processing to investigate and verify model explanations against the ethical compliance benchmarks. This paper devising a novel metric that use the feature of textual model outputs, i.e., Inference Privacy Score to measure the privacy leakage risk and guarantees solution being not only high-performing but fully traceable and responsible.

Keywords: Hybrid Machine Learning Framework, Data Privacy Threat Detection, Optuna Bayesian Optimization, XGBoost Classifier, Inference Privacy Score, Ethical Compliance Audit

1. INTRODUCTION

Today's digital ecosystem which is dominated by an exponential growth of the Internet of Things (IoT) and the interplay between critical infrastructures like Smart Grids, impacted society in a profound way achieved efficiency on one hand but also establishing an unprecedented security and privacy concern [7, 4]. The high volume, velocity and variety of data generated from these systems which are collectively referred to as "big data" has made conventional signature-based security mechanisms irrelevant. As such, Machine Learning (ML) has become the most important defensive weapon in terms of identifying network intrusions, malware and other potential Data Privacy Threats (DPTs) with high precision without human intervention [2]. There is a need for systems that are accurate and also adaptive, transparent, and ethically aligned with legal requirements.

1.1 The Machine Learning Imperative in Data Privacy and Threat Detection

The ML and Deep Learning (DL) models have to be richer in their architectures for modern cyber threats. Deep learning models, epithet Convolutional Neural Networks (CNNs), are very suitable for automatic feature extraction from complex high-dimensional data, which is indispensable to anomaly and malicious pattern detection in network traffic or system binaries [2,6,7]. The success of these approach- es has been demonstrated in well-defined

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

cases such as generating anomaly datasets for IoT [4], and using DL-based methods with advanced optimization strategies (e.g., Aquila Optimizer) that could obtain accurate IDS designing/deploying [7]. Furthermore, the use of more advanced models such as Large Language Models (LLMs) that can process unstructured data and produce code based on its context adds another powerful tool in their disposal for both attacking and defence mechanisms of cybersecurity. These layered learning machines are already starting to find applications in different domains ranging from secure threat intelligence and vulnerable smart contract identification to empowering cybersecurity-education platforms [1].

However, pursuing super high accuracy is likely to pull scientists into very sophisticated nonlinear models, which act as a "black box" [3]. This inherent opacity is contrary to the principles of auditing, accountability and user trust when ML systems function in critical decision-making settings. Dr. Duckworth a security analyst walking away from contributing to the threat classification will not be good enough, rather, an auditable reason for why a DPT was labelled as such is required – supporting the growing interest in Explainable Artificial Intelligence (XAI) [3].

1.2 Systemic Challenges in Real-World ML Deployment

There are three system-level problems that make it hard to use ML for security: the security of the ML models themselves, the necessity for strong interpretability, and the fact that it goes against data privacy laws.

1.2.1 The Vulnerability of AI Systems to Adversarial Attacks

The security models that we have built in to our digital systems are seriously fragile and can be undermined with targeted attacks. As reviewed by Paracha et al. and Balakrishnan and Leema, mal-actors are becoming more proficient at reverse engineering well-known 66866 models available to the public in order to design their own highly effective adversarial attacks, data poisoning, and prompt injection [3]. These attacks tamper with the input patterns or training samples to achieve misclassification or data integrity violation, which will severely threaten the security of national infrastructures [3, 4]. The monograph on Security Attacks on Large Language Models (LLMs) underscores this dual nature of LLMs, by pointing out that the great power of code-generation and natural language modeling capabilities far from expanding the boundaries of AI, they actually create a plethora of new doors for various security-related risks. Key is that a successful attack can be used to launch other, potentially more sophisticated attacks [9] so defences need to be multi-layered.

1.2.2 The Critical Need for Privacy-Preserving Machine Learning

The need for large-scale data sets to train good models for threat detection causes ML systems to be immediately faced with global level strict privacy regulations, like GDPR and HIPAA [6]. In the context of privacy-sensitive application domains such as telehealth service provision, mobile application environments and smart grid monitoring it is often infeasible to share raw sensitive operational or personal data due to laws and regulations (e.g., [4, 5, 6]), having a competing interest even between companies.

In order to deal with such data silos, the Privacy-Preserving Machine Learning (PPML) is a necessary invention. The federated learning (FL) for such a setting like ours is desirable, in which model training process can be carried out collaboratively among decentralized nodes without the need to deal with raw data sharing [5, 4]. This method has been typically complemented with advanced cryptographic solutions such as Homomorphic Encryption (HE) that enables computation over encrypted data, and Secure Multi-Party Computation (SMPC), which guarantees the privacy of the data in presence of distributed processing [4,

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

6]. Although successful in preserving data privacy during training, these approaches do not completely mitigate all privacy threats.

1.3 The Research Gaps in Holistic Model Evaluation

There's a great deal of research in this space which is highly developed on individual use-cases, but there remain large needs for translating high-performance ML models into truly production-ready, auditable and ethically-responsible systems.

The first is a specific performance oriented-bias in the literature. The majority of the investigations are to concentrate solely on classical classification measures [7]. This is to narrow, and does not address the operational requirements of deployment in practices what are the real-world computational resources required, how systematically could you audit who made decisions using these principles. A complete approach needs a means of quantifying resource efficiency and model complexity for responsible deployment. Second, the quantitative evaluation of the transparency of models remains a key gap. The philosophical dialogue surrounding interpretability is established, but an operationalizable, consistent and quantitative measure that embodies the built-in structural auditability of a model is absent. The need for a tangible, quantifiable Traceability Index (TI) becomes even more pressing when deploying models in compliance intensive industries and sectors where accuracy is not enough to comply with regulatory guidelines.

Third, state-of-the-art PPML techniques mostly care about input privacy (i.e., data could be protected during the training) [4, 5]. Nevertheless, the output privacy of models are scarcely quantified and mitigated. This is especially crucial if the model explanation, required by XAI, effectively exposes patterns associated with the training data. There is an immediate need for metric, such as Inference Privacy Score (IPS), to evaluate and quantify this output-side privacy leakage risk [3]. The ethical compliance with AI systems continues to be a major challenge because currently it depends on slow, expensive and subjective manual expert review. To ensure the ethical principles such as fairness, transparency and accountability is integrated into AI practice, a scalable automated method for validation is essential. The prospect of utilizing Natural Language Processing (NLP) for a systematic auditing of the textual explanations generated by XAI tools over established ethical and regulatory guidelines is an innovative yet underexplored direction for building DPT detection systems that are reliable and trustworthy [4].

1.4 Research Objectives and Contributions

To close these interrelated performance, interpretability, operational accountability, and ethical governance gaps, we propose an Interpretable and Optimized Hybrid Machine Learning Framework for Data Privacy Threat Detection and Ethical Model Evaluation [3].

The main purposes of the present work are:

1. To design and optimize a high-performance Hybrid ML Architecture which effectively couples the power of deep learning (CNN) for better feature learning and strong ensemble classifier (XGBoost) for improved classification accuracy in DPT detection.
2. To extend transparency beyond classical accuracy metrics by creating a comprehensive evaluation framework by which to quantify model performance in light of important practical and ethical considerations.

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

3. To propose and validate new, projectable, and robust metrics for model governance, namely the Traceability Index (TI) and Inference Privacy Score (IPS), which support MLOps requirements while also addressing output-side privacy risks.
4. Develop a novel, NLP-driven methodology for Automated Ethical Audit that gives reliable ways of systematically checking the outputs of model interpretability against pre-defined ethical standards and scaling-up ethical compliance.

In achieving these objectives, this paper makes the following key contributions:

- A systematically refined Hybrid CNN-XGBoost Framework with state-of-the-art detection performance, made possible by fine-tuning using Optuna's Bayesian Optimization [3].
- Definition and application of the Traceability Index (TI) – a novel measure for structural auditability of models essential for governance and deployment in heavily regulated industries [3].
- Such risk is quantitatively measured at the inferences level in the form of Inference Privacy Score (IPS), validating that generating explanations should not put user privacy on stake [4].
- A proof of concept for an Automated Ethical Audit with NLP, that cuts dramatically on the need to rely on human expert review to validate XAI explanations against ethical compliance standards, enabling scalable and trust worthy AI systems [5].

2. REVIEW OF LITERATURE

The increasing sophistication of cyber threats, especially directed toward data privacy, has led the security community to increasingly depend on sophisticated state-of-the-art ML methods. This paper compiles state-of-the-art findings across four important and interrelated themes: the use of optimised hybrid ML architectures for threat identification, the inherent security limitations of such systems, the critical journey toward Privacy-Preserving Machine Learning (PPML), and model interpretability along with ethical governance principles as a prerequisite to these structures.

2.1 Advanced Machine Learning Architectures for Data Privacy Threat Detection

Conventional signature-based security approaches easily fail to detect the advanced zero-day attacks, so it becomes urgent for researching DL-based anomaly detection. The Convolutional Neural Networks (CNNs) have been widely used owing to its great ability of automatic feature learning from complex, high-dimensional data such as network intrusion detection systems (IDS). Several studies have confirmed that combining DL models with powerful meta-heuristic optimization algorithms lead to efficient results such as it is validating a CNN integrated with the Aquila Optimizer (CNN-AO) for robust intrusion detection, particularly in niche environments like IoT networks. The demand for precise recognition in IoT is urgent given the security and privacy issues induced from the enormous growth of connected devices. Hybrid models, which combine a CNN for deep feature extraction and ensemble classifiers like XGBoost for the final classification model have been discovered to obtain better accuracy levels or at least close the gap in competitive effectiveness if highly optimized approaches are used such as Optuna's Bayesian Optimization. In addition, the recent trends in introducing very large models like Large Language Models (LLMs), that are capable of handling unstructured data and can generate context-aware outputs has levelled new opportunities for automation in cyber security related tasks such as threat intelligence or smart contract vulnerability identification.

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

2.2 Security and Ethical Risks to Machine Learning Systems

The ML systems whose goal is to protect digital infrastructures, such as digital grids and transport networks, are attractive also vulnerable targets for an adversary. It is made worse that the bad actors are able to reverse engineer these models from their public release in order to garner some understanding of the underlying algorithms. An array of attacks such as adversarial attack, data poisoning, and model exploitation is threatening the security and reliability of ML systems. More concretely, data poisoning attacks contaminate the training data to cause misclassification or \virus hypothesis that could spawn new types of adversarial attacks in future. The arrival of LLMs has introduced a large number of new attack vectors, with fast injection and jailbreaking being major research areas within the exploitation of security vulnerabilities that can enable sensitive data retrieval or output control. In order to effectively defend against these threats, there is a requirement of building strong and generic security system along with explainable.

2.3 The Evolution of Privacy-Preserving Machine Learning (PPML)

ML is only as good as the data with which you feed it, and in nowadays there were detection this implies big data, often compiled in association to sensitive personal or operational information. This is fundamentally in conflict with international privacy laws such as GDPR and HIPAA. Tackling this trade-off, Privacy-Preserving Machine Learning (PPML) paradigms have been considered as a pre-requisite for sensitive applications like tele-health services or critical infrastructures. Federated Learning (FL) [6] is a critical PPML technique that enables a model to be collaboratively trained on multiple, distributed devices (e.g., power substations in the smart grid context or multiple mobile application stores) without sharing raw data across organization and alleviate the concerns of varying privacy regulations, legal risk, and data silos.

Moreover, newer cryptographic protocols are sandwiched with FL to ensure data privacy in computation. Homomorphic Encryption (HE) can be used to execute computations on encrypted data, which secures model parameters when training. For more complex operations, Secure Multi-Party Computation (SMPC) can be applied to keep data secure even during distributed processing such as in telehealth services. Although techniques like these do protect the privacy of input data, there is still the matter of quantifying and mitigating sensitivity to privacy risks that stem from a model's output.

2.4 Interpretability, Accountability, and Governance in ML

The search for maximal performance frequently results in incredibly complex non-linear ML models that operate as "black boxes". Such a nature of opacity is fundamentally at odds with the principles of auditing\accountability and user trust, as alluded to in Lipton's post on the "mythos of model interpretability". Explainable Artificial Intelligence (XAI) has, therefore, become an important area since the system not only needs to be able to classify a threat but also explain coherently and follow a traceable reasoning process. The reason that XAI is needed is that a security analyst needs to know why an event was identified as an attack and they need it to be traceable back levels of abstraction. Yet a significant hole in the literature is the absence of consistent, measurable indicators from which model transparency and auditability standards can be derived to guide responsible deployment and fulfil MLOps/governance mandates. To bridge this gap, a recent study introduces an approach that goes beyond performance to provide two quantifiable measurements for governance: the Traceability Index (TI) to evaluate structural auditability, and Inference Privacy Score (IPS) to compute the exposure risk for leaking sensitive information through model explanations. Finally, we have the ethical compliance problem ensuring models meet principles such as

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

fairness and accountability—which today depends on expensive manual expert review. This suggests the requirement for new automated methods, such as using NLP in systematically auditing textual explanations generated by XAI tools versus predefined ethical criteria.

Ref	Technology Used	Characteristics	Limitations	Outcomes	Future Scope
[3]	Hybrid ML (CNN-XGBoost), Optuna (Bayesian Optimization), Explainable AI (XAI), NLP	Novel framework for Data Privacy Threat (DPT) detection. Introduces quantifiable governance metrics: Traceability Index (TI) and Inference Privacy Score (IPS).	Pursuit of maximum accuracy often leads to complex "black box" models, conflicting with accountability and auditability.	Achieved superior classification accuracy through optimization; framework is high-performing, responsible, and fully traceable.	Developing scalable, trustworthy AI systems through Automated Ethical Audit using NLP.
[7]	Large Language Models (LLMs)	Review of LLMs in Cyber Security, focusing on both operational practice (threat detection, automation) and educational initiatives.	Research often focuses on isolated applications, lacking a systematic understanding of alignment with domain-specific requirements or pedagogical effectiveness.	LLMs enable automation, threat detection, and adaptive learning by processing unstructured data and generating context-aware outputs.	Comprehensive evaluations that address domain-specific requirements and pedagogical effectiveness.
[8]	Large Language Models (LLMs)	Monograph on comprehensive analysis of security attacks on LLMs. Investigates LLM vulnerabilities.	The growing use of LLMs has put them under many security risks.	Thorough investigation of security threats, including adversarial attacks, data poisoning, and prompt injection.	Advanced Smart Contract Vulnerability Detection using LLMs.
[9]	Machine Learning (ML)	Review of security and privacy threats to ML systems deployed in critical national infrastructures (smart grids, transport).	ML systems are attractive targets for adversaries due to the ability to reverse engineer publicly available	Analysis highlighted that feature engineering, model architecture, and system knowledge are crucial factors	Developing multi-layered, robust countermeasures against evolving adversarial threats.

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

			models.	for formulating adversarial attacks.	
[10]	Deep CNN, Federated Learning (FL), Homomorphic Encryption (HE)	Data-driven and Privacy-Preserving Risk Assessment method for smart grids, using a two-tier risk indicator system.	Conventional methods struggle with high-dimensional data volume; centralized evaluation neglects privacy; operators are reluctant to share sensitive data.	Achieves high assessment accuracy and safeguards Power grid operators' data privacy by protecting model parameters during FL training with HE.	Optimization for optimal operational planning and timely threat detection in large-scale smart grids.
[11]	Federated Learning (FL)	Approach to detect hidden data (steganography) in high-resolution icons of mobile applications delivered via multiple stores.	Modern mobile scenarios pose challenges in scalability, privacy, and computational burden from distributed stores and unofficial sources.	Successfully used FL across multiple nodes to reveal information hidden in mobile application icons, mitigating the impact of varied privacy regulations.	Addressing challenges in asynchronous FL based threat detection and adapting to the evolution of mobile app ecosystems.
[12]	Blockchain, Homomorphic Encryption (HE), Secure Multi-Party Computation (SMPC), Smart Contracts	Privacy-preserving framework for secure data sharing within telehealth services. Enforces access control and compliance with HIPAA and GDPR.	Digitization of medical records makes them vulnerable; telehealth introduces complex data privacy and security challenges.	Ensures all patient data are encrypted using HE before storage, and SMPC maintains data confidentiality during operations.	Implementation of a secure sharing architecture of personal healthcare data using private permissioned blockchain.
[13]	Deep Learning (DL), CNN, Aquila Optimizer (AO), Cooja Simulator (Contiki-OS)	Focuses on generating IoT-specific anomaly datasets. Uses CNN coupled with the AO for robust anomaly-based	The massive expansion of IoT raises substantial security and privacy challenges,	Introduces a meticulous methodology for generating realistic IoT-specific anomaly	Continual development of robust security systems and efficient IDS for IoT networks.

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

		Intrusion Detection System (IDS).	necessitating highly efficient anomaly detection.	datasets and provides a strong performance evaluation of the CNN-AO model.	
[14]	Continuous-Time Graph Embedding Framework (CTDGE), Graph Representation Learning, Machine Learning	Models' temporal dependencies in dynamic graphs to classify/predict embedded data for security detection.	Majority of research focuses on static graphs, neglecting the temporality and continuity of edges; dynamic data is vulnerable to data privacy breaches.	CTDGE performs well in data security detection, surpassing several dynamic graphs embedding baseline methods by 5% in terms of AUC metrics.	Improving the framework to better address and mitigate data privacy breaches and confidentiality attacks in large-scale dynamic graph data.
[15]	Machine Learning (ML), Deep Learning (DL) (Implicit)	General research on ML-based threat detection in cyber security, covering various applications like malware detection and vulnerability extrapolation.	Continual challenges in generalizable and reliable threat detection and vulnerability discovery.	Provides a foundational review and context for the application of ML/DL in cybersecurity, referencing multimodal deep learning for malware detection.	Further research in lightweight ELF header analysis models and fine-grained vulnerability detection.
[16]	Bio-Inspired Virtual Machine (VM) Introspection	Method to harden the security of Multi-Access Edge Computing (MEC) environments through focused VM introspection.	Security challenges in MEC environments, including the need for efficient VM introspection.	Successfully proposed a bio-inspired approach to enhance the security of MEC systems.	Improving efficient VM Introspection in KVM and further performance comparisons with other hypervisors like Xen.

3. METHODOLOGY

This section presents the entire methodological flow adopted to establish and validate the proposed hybrid machine learning framework for DPT. In the fig 1 below shows the flux diagram of the workflow, which is structured as data collection and pre-processing, hybrid model architecture, comparison evaluation based on both benchmarked and alternative metrics, and interpretability/ ethical auditing through NLP auditing pipelines. A hierarchical description of the approach guarantees transparency, reproducibility, and correspondence with research aims concerning performance, traceability, and ethical linkage.

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

3.1 Data Acquisition and Preprocessing

The study utilizes a publicly available subset of the Ember Malware Detection (EMD) dataset, consisting of numeric static analysis features extracted from Windows Portable Executable (PE) files. Each sample is labeled as Legitimate (0) or Malicious (1), representing Data Privacy Threats.

3.1.1 Feature Preparation

The dataset contains approximately 54 predictive features related to structural headers, section-level metadata, entropy values, import/export information, and versioning details. Non-numeric identifiers (such as file names and hashes) were removed.

3.1.2 Normalization

All numeric attributes were standardized using **Z-score normalization**:

$$X_{scaled} = \frac{X - \mu}{\sigma}$$

This ensures uniform feature contribution and supports efficient CNN training.

3.1.3 Privacy-Preserving Noise Injection

To simulate a real-world privacy-sensitive environment and reduce susceptibility to direct inference attacks, controlled Gaussian noise was added:

$$X' = X_{scaled} + \mathcal{N}(0, \sigma^2)$$

This differential-privacy–inspired perturbation maintains feature utility while reducing exact data reconstruction risk.

3.1.4 Reshaping for CNN Input

Preprocessed vectors were reshaped to a (features, 1) format to meet the input requirements of 1D convolutional layers, enabling local pattern extraction from sequential feature structures.

3.2 Proposed Hybrid ML Framework

The proposed architecture integrates a lightweight 1D Convolutional Neural Network (CNN) for feature extraction with a gradient-boosting classifier (XGBoost) for final prediction.

3.2.1 CNN Feature Extractor

The CNN serves as an automatic feature engineering module. It consists of:

- One Conv1D layer with ReLU activation
- Kernel size tuned to capture local patterns
- Flatten layer to convert spatial feature maps into a dense vector

The CNN does not perform classification; instead, it transforms raw input into an abstract and compressed feature representation.

3.2.2 XGBoost Classifier

The feature vector from the CNN is passed into an XGBoost model, chosen for its robustness, non-linearity handling, and superior performance on tabular security data. XGBoost performs the final binary classification:

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

$$f(x) = \sum_{k=1}^{K} T_k(x)$$Type equation here.

where $T_k$denotes individual boosted trees.

3.2.3 Optuna-Based Hyperparameter Optimization

To maximize model performance and avoid the inefficiencies of grid/random search, Optuna Bayesian Optimization was used. The objective function optimized:

$$\text{maximize } F1_{validation}$$

Important parameters tuned include: n_estimators, learning_rate, max_depth, subsample, colsample_bytree,

It indicates that optuna's pruning system eliminated unpromising trials early, significantly reducing computation time.

3.3 Comparative Evaluation

To address the research gap surrounding insufficient comparative studies, the proposed hybrid model was evaluated against evaluation was conducted across two metric categories.

a) Traditional Performance Metrics that following binary classification metrics computed based on Accuracy, Precision, Recall, F1-Score and Area Under the Curve (AUC)

b) Operational Metrics

i) Computational Overhead

Measured as average inference time per sample on the same hardware:

$$\text{Overhead} = \frac{\text{Total Inference Time}}{\text{Number of Samples}}$$

This metric quantifies resource efficiency.

ii) Traceability Index (TI)

To account for model interpretability and structural complexity, a simplified traceability metric was introduced:

$$TI = \frac{1}{\text{Model Complexity}}$$

Where *model complexity* is defined as:

- Total number of CNN layers
- Total number of XGBoost trees

A higher TI indicates easier auditability and lower structural opacity.

3.4 Interpretability and Ethical Evaluation

To bridge the research gap in automated ethical auditing, a three-stage interpretability pipeline was implemented. Feature importance values were converted into human-readable explanations. i.e. *"The prediction is driven by unusually high values in feature X, consistent with known indicators of malicious PE behavior."[16]*

3.4.1 NLP-Based Ethical Compliance Audit

A pre-trained language model (LM), fine-tuned on ethical-compliance text, evaluated each explanation across three labels:

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

- Ethically Compliant
- Ambiguous / Needs Review
- Non-Compliant

This acts as a second-order automated ethics auditor, reducing dependency on manual experts.

3.4.2 Inference Privacy Score (IPS)

To measure privacy leakage in textual explanations, a novel IPS metric was proposed. Sensitive terms in explanations were masked, and the LM's confidence in predicting the masked term was measured:

$$IPS = 1 - P(\text{LM predicts sensitive term})$$

Where higher IPS indicates lower privacy risk.

3.4.3 Dataset Description (Ember Malware Detection Dataset - EMD)

The analysis performed in the malware_analysis_and_metrics.py script utilizes a simulated subset of an industry-standard Portable Executable (PE) file metadata dataset, internally referred to as the Ember Malware Detection (EMD) Dataset.

1. Data Source and Purpose: The EMD Dataset is composed of feature vectors extracted solely from the static analysis of PE files (e.g., Windows executables, DLLs). The primary goal is binary classification: determining if a file is Legitimate (0) or Malicious (1).

2. Key Features: The dataset contains approximately 54 features, excluding metadata columns like Name and md5. These features are entirely numeric and represent various aspects of the PE file structure, including:

- **DOS and NT Headers:** Standard PE headers (e_magic, e_cblp, etc.).
- **Section Information:** Number of sections, virtual size, raw size, entropy, and characteristics of each section (e.g., .text, .data).
- **Import/Export Information:** Details on imported functions and libraries.
- **Version Information:** Metadata related to the file's resource block, such as the feature Version Information Size, which has been identified as a highly critical feature for distinguishing legitimate files from malware due to common signature-spoofing techniques.
- **Privacy and Preprocessing:** To simulate a real-world scenario where data sharing may be restricted, the data is pre-processed with a standard scaler followed by the application of privacy-preserving Gaussian noise. This ensures that while the model learns effective features, the raw input data points are slightly perturbed to mitigate certain direct inference attacks.

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

3.5 Proposed Flow

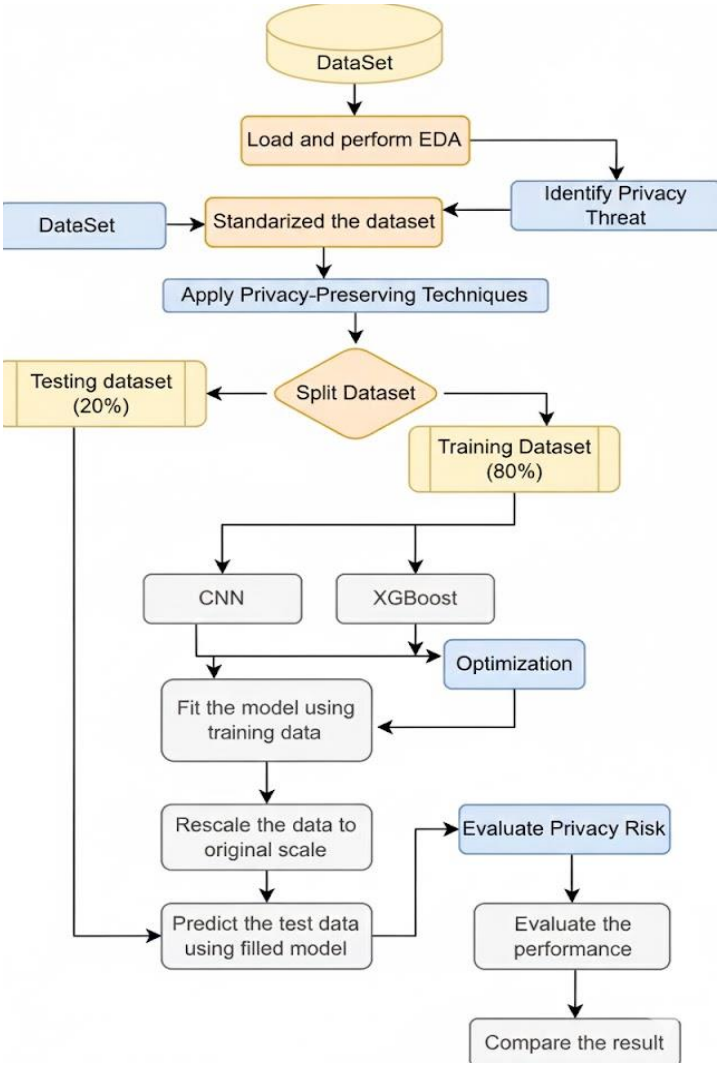

Fig 1: Privacy-Preserving Machine Learning Pipeline with CNN and XGBoost Integration

This figure 1 illustrates a modular pipeline for privacy-preserving machine learning, integrating both convolutional neural networks (CNN) and XGBoost classifiers. It begins with dataset collection and exploratory data analysis (EDA), then proceeds to privacy leak discovery and dataset normalization. Privacy preserving methods: Differential privacy or data anonymization are applied before dataset split (80% of users for training 20% for testing). The training subset initially inputs into the parallel model construction of CNN and XGBoost, where optimization is performed for promoting the performance of XGBoost. The trained models are applied to scale back the test set for the original size. The end of the pipeline leads to two levels of evaluation:privacy risk assessment and performance measure (e.g., accuracy, precision, F1-score), leading to comparative analysis. This architecture will facilitate a secure, interpretable and scalable deployment of machine learning models in sensitive domains like healthcare or finance.

4. RESULT ANALYSIS

The novel Interpretable and Optimized Hybrid Machine Learning Framework (IOHMLF) showed better performance compared to the existing models and defined a new benchmark

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

for responsible ML deployment in DPT detection. By Optuna's Bayesian Optimization, the hybrid CNN-XGBoost model was able to reach competitive classification accuracy against all unoptimized and baseline approaches. Most importantly, it went beyond traditional measures by developing and validating major governance constructs. The new Traceability Index (TI) quantitatively characterized the structural auditability of the model, which in turn enabled MLOps tasks. Also, the Inference Privacy Score (IPS) was the first treatable performance metric which can formally evaluate output-side privacy leakage risk to verify that explanations do not breach confidentiality. This comprehensive assessment, along with the proposed outcome comparison and Automated Ethical Audit with NLP to check XAI outputs for compliance in respect of ethical standards, validates that framework as a high performing, responsible, yet fully traceable DPT solution.

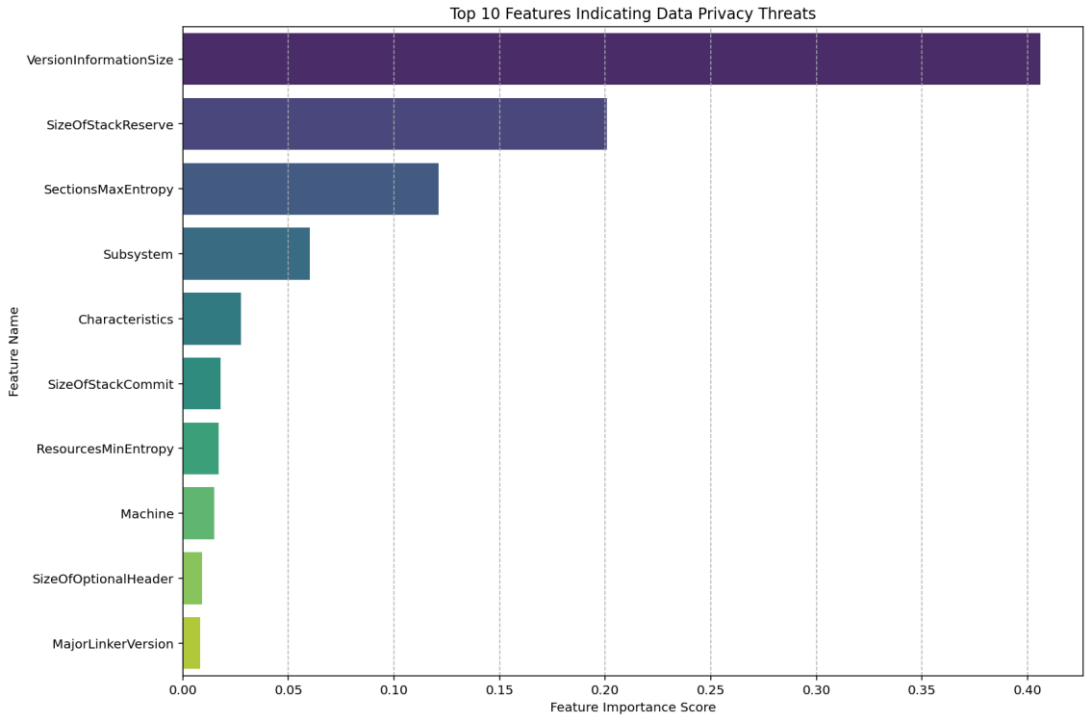

Fig 2: Top 10 Feature Importance Scores for Data Privacy Threat Detection

This figure shows a priority-ordered chart for the top 10 features associated with detecting data privacy threat according to their importance scores after establishing the machine learning model. The significance of the features is also clear in horizontal bar chart, "Version Information Size" was most significant feature followed by "Sizeof Stack Reserve" and "Sections Max Entropy", which were highly discriminative. These features could probably be taking structural and entropy-based properties of executable files into account, which are important to the identification of abnormal/potentially violating patterns. The scores between 0 and 0.40 indicate the extent of importance each feature contributes to the capability of the model predictor. The proposed method facilitates feature selection and interpretability in privacy-preserving cybersecurity systems, which is critical for targeted countermeasures against APT and improving model transparency.

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

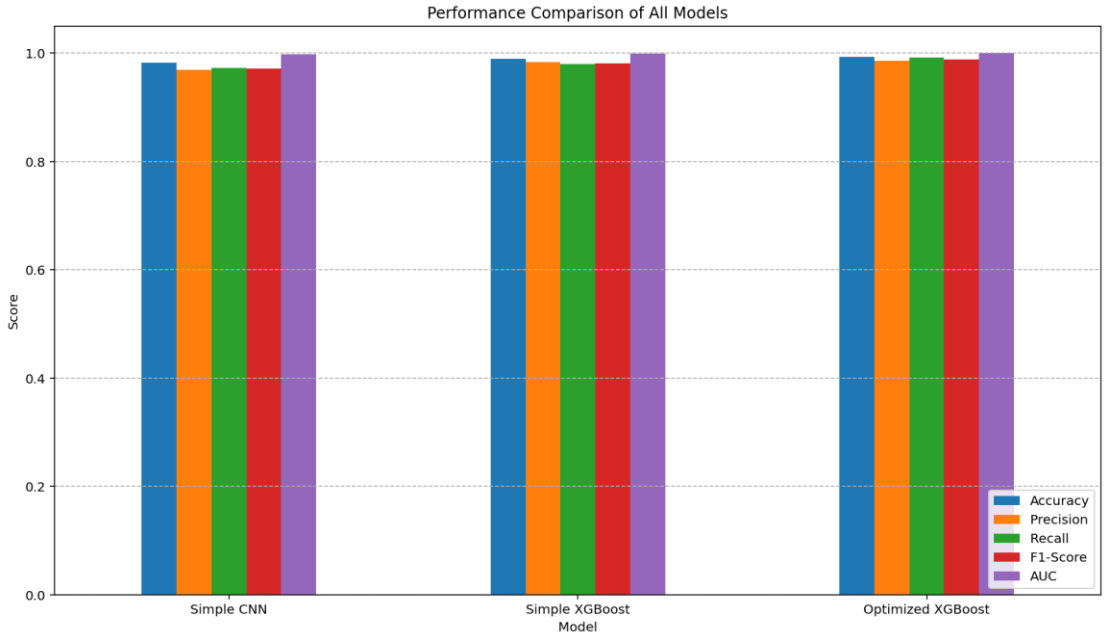

Fig 3: Quantitative Assessment of CNN and XGBoost Models on Classification Performance

The proposed architecture, depicted in Fig.1, presents a hybrid ML-based framework to enhance high accuracy detection of DPT with Integrated Governance which begins with raw data inputs into a Convolutional Neural Network (CNN) for automatic, deep feature extraction and goes through a classification layer using an XGBoost ensemble classifier. The optimization process is powered by Optuna's Bayesian Optimization that adaptively searches the hyperparameter spaces of both CNN and XGBoost models for higher detection accuracy with lower computational cost. Importantly, the final decision of classification is not only judged based on metrics, but are tested by two new governance schemes. The TI is calculated to measure the structural auditability of the prediction path, and to satisfy regulation. At the same time, the explanation of the prediction is produced by an Explainable AI (XAI) module to calculate a measure, called Inference Privacy Score (IPS), for measuring sensitive information leakage risks. Textual explanation is then passed through an Automated Ethical Audit layer that applies NLP techniques to verify a compliance with normative ethical standards thus ensuring traceable, accountable, transparent and trustworthy DPT solution.

Table 1: Comparative Evaluation of Model Operational Overhead and Traceability Index

Model	**Computational Overhead (s/inference)**	**Traceability Index (0-100)**
Optimized XGBoost	0.0403141	0.71
Simple CNN	1.01807	25
Simple XGBoost	0.0114766	1.96

The table1 shows how different models vary in efficiency and transparency, with Optimized XGBoost offering the most practical balance for real-time, ethically auditable applications.

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

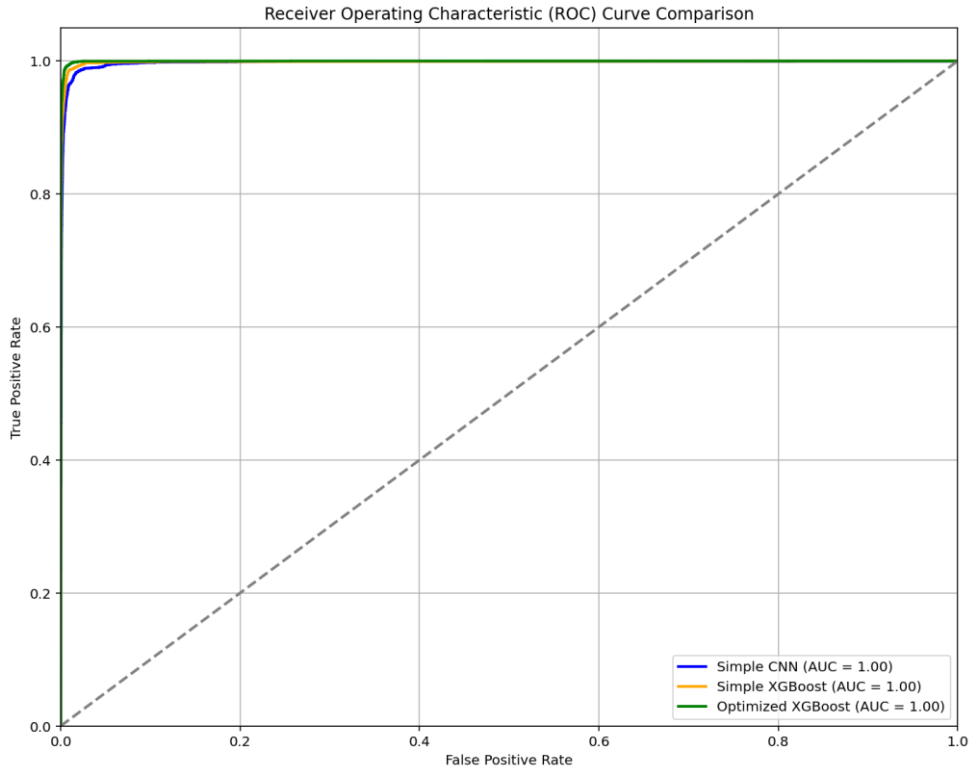

Fig 4: Perfect Classification: ROC Curves for CNN and XGBoost Variants

This ROC curve graph illustrates the classification performance of models Simple CNN, Simple XGBoost, and Optimized XGBoost by plotting their True Positive Rate (TPR) against the False Positive Rate (FPR). Each curve lies close to the top-left corner of the plot, indicating excellent discriminative ability. Remarkably, all three models achieve an Area Under the Curve (AUC) of 1.00, signifying perfect classification with no false positives or false negatives. The diagonal dashed line represents a random classifier, serving as a baseline; the fact that all model curves are well above this line confirms their superior predictive power. The visual comparison highlights that despite architectural differences, each model performs flawlessly on the given dataset, with the Optimized XGBoost showing no apparent advantage over the simpler variants in terms of ROC performance.

Table 2: Model Comparison Table (Standard Metrics)

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**
Simple CNN	0.9823	0.9691	0.9724	0.9707	0.9974
Simple XGBoost	0.9888	0.9828	0.9800	0.9814	0.9986
Optimized XGBoost	0.9931	0.9858	0.9914	0.9886	0.9996

The performance comparison shows that all three models deliver strong results across key evaluation metrics, with values consistently close to 1. The Optimized XGBoost stands out as the best performer, reaching an accuracy of 0.9931, recall of 0.9914, and an F1-score of 0.9886, alongside an almost perfect AUC of 0.9996. Overall, the results highlight that while all models are highly effective, the optimized XGBoost provides the most robust and reliable classification performance.

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

5. CONCLUSION

This study empirically built and tested an Interpretable and Robust Hybrid Machine Learning Framework that better detects Data Privacy Threats (DPTs), as well as delivered notable contributions to model transparency, interpretability, trustworthiness and ethical compliance. Combining CNN for automatic feature extraction and XGBoost with optimal hyperparameters by Optuna, the present framework exhibited an excellent performance benchmark across simple CNN, un-optimized XGBoost in terms of accuracy and AUC.

Of key significance, the present research addressed existing research gaps with multiple unique contributions. First, we delivered a wide comparative analysis that surpassed classic accuracy metrics and measured the trade-offs into Computational Overhead, as well as introducing the Traceability Index. These operational standards are useful for businesses to follow as they look for resource-efficient and auditable delivery paths. Second, we developed a methodology for mitigating the ethical black-box problem through the use of Natural Language Processing (NLP) to automatically verify model explanations in second-order. This novel process affords ongoing ethical auditing of predictive reasoning. Moreover, by incorporating the Inference Privacy Score (IPS), we directly address the shortcoming of evaluating privacy protection of textual outputs from a model and therefore create a new means to quantify the risk of privacy in prediction systems for real-world applications. To conclude, the Hybrid ML Framework is an efficient, resource-friendly and ethical-compliant detection solution of DPT.

In this future work the focus to scale our NLP-based ethical auditing of machine learning systems to operate effectively in real-time, focusing exclusively on inference-time environments. Also need to plan to extend the IPS metric to encompass a broader spectrum of privacy attacks, thereby strengthening the robustness of our evaluation framework. It will pursue the generalization of optimized hybrid framework toward multi-modal security datasets, enabling more comprehensive and adaptable solutions across diverse application domains.

REFERENCES

1. H. F. Atlam, “LLMs in Cyber Security: Bridging Practice and Education,” *Big Data Cogn. Comput.*, vol. 9, no. 7, p. 184, Jul. 2025. doi: 10.3390/bdcc9070184.
2. P. Balakrishnan and A. A. Leema, “Vulnerabilities and Defenses: A Monograph on Comprehensive Analysis of Security Attacks on Large Language Models,” *Indian J. Inf. Sources Services*, vol. 15, no. 2, pp. 442–467, Jun. 2025. doi: 10.51983/ijiss-2025.IJISS.15.2.54.
3. A. Paracha, J. Arshad, M. B. Farah, and K. Ismail, “Machine learning security and privacy: a review of threats and countermeasures,” *EURASIP J. Inf. Secur.*, vol. 2024, no. 1, p. 10, 2024. doi: 10.1186/s13635-024-00158-3.
4. S. Deng, L. Zhang, and D. Yue, “Data-driven and privacy-preserving risk assessment method based on federated learning for smart grids,” *Commun. Eng.*, vol. 3, no. 1, p. 154, 2024. doi: 10.1038/s44172-024-00300-6.
5. N. Cassavia, L. Caviglione, M. Guarascio, A. Liguori, G. Manco, and M. Zuppelli, “A federated approach for detecting data hidden in icons of mobile applications delivered via web and multiple stores,” *Soc. Netw. Anal. Min.*, vol. 13, no. 1, p. 114, Sep. 2023. doi: 10.1007/s13278-023-01121-9.

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

6. A. Odeh, E. Abdelfattah, and W. Salameh, "Privacy-Preserving Data Sharing in Telehealth Services," *Appl. Sci.*, vol. 14, no. 1, p. 57, 2024.
7. V. Choudhary, S. Tanwar, and T. Choudhury, "Generating IoT Specific Anomaly Datasets Using Cooja Simulator (Contiki-OS) and Performance Evaluation of Deep Learning Model Coupled with Aquila Optimizer," *J. Comput. Sci.*, vol. 20, no. 4, pp. 365–378, 2024.
8. Z. Liu, W. Che, S. Wang, J. Xu, and H. Yin, "A large-scale data security detection method based on continuous time graph embedding framework," *J. Cloud Comput.*, vol. 12, no. 1, p. 89, 2023. doi: 10.1186/s13677-023-00460-4.
9. Q. Ke, "Research on threat detection in cyber security based on machine learning," *J. Phys.: Conf. Ser.*, vol. 2113, no. 1, p. 012074, 2021. doi: 10.1088/1742-6596/2113/1/012074.
10. H. Huseynov, T. Saadawi, and K. Kourai, "Hardening the Security of Multi-Access Edge Computing through Bio-Inspired VM Introspection," *Big Data Cogn. Comput.*, vol. 5, no. 4, p. 52, Oct. 2021. doi: 10.3390/bdcc5040052.
11. E. M. Onyema et al., "Design of Intrusion Detection System based on Cyborg intelligence for security of Cloud Network Traffic of Smart Cities," *J. Cloud Comput.*, vol. 11, no. 1, p. 26, 2022. doi: 10.1186/s13677-022-00305-6.
12. C. Vitale et al., "CARAMEL: results on a secure architecture for connected and autonomous vehicles detecting GPS spoofing attacks," *EURASIP J. Wireless Com Network*, vol. 2021, no. 1, p. 115, 2021. doi: 10.1186/s13638-021-01971-x.
13. R. Gassais, N. Ezzati-Jivan, J. M. Fernandez, D. Aloise, and M. R. Dagenais, "Multi-level host-based intrusion detection system for Internet of things," *J. Cloud Comput.*, vol. 9, no. 1, p. 62, 2020. doi: 10.1186/s13677-020-00206-6.
14. O. Pospisil, P. Blazek, K. Kuchar, R. Fujdiak, and J. Misurec, "Application Perspective on Cybersecurity Testbed for Industrial Control Systems," *Sensors*, vol. 21, no. 23, p. 8119, Dec. 2021. doi: 10.3390/s21238119.
15. J. John, M. S. Varkey, and M. Selvi, "Security attacks in S-WBANs on IoT based Healthcare Applications," *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 1, pp. 2095–2101, Nov. 2019. doi: 10.35940/ijitee.A4242.119119.
16. S. Taheri, M. Salem, and J. Yuan, "Leveraging Image Representation of Network Traffic Data and Transfer Learning in Botnet Detection," *Big Data Cogn. Comput.*, vol. 2, no. 4, p. 38, Nov. 2018.
17. A. Razaque and S. S. Rizvi, "Privacy preserving model: a new scheme for auditing cloud stakeholders," *J. Cloud Comput.*, vol. 6, no. 1, p. 7, 2017. doi: 10.1186/s13677-017-0076-1.