Ghulam Ishaq Khan Institute of Engineering Sciences and Technology (GIKI)

Final Report

for

PhishGuard: An Explainable AI Approach for Email Threat Detection

Version 1.0

by

Member 1 Muhammad Shaheer	**2023509**
Member 2 Muhammad Bin Waseem	**2023403**
Member 3 Muhammad Zaeem Nawaz	**2023550**
Member 4 Abdullah Khan Mahsud	**2023346**

Supervisor

Mr. Ahmed Nawaz

Bachelor of Science in Cyber Security (2023–2027)

Faculty of Computer Science and Engineering (FCSE)

Table of Contents

Contents

1. Abstract

Phishing emails remain one of the most prevalent vectors for cyberattacks, exploiting social engineering and deceptive content to compromise users. Traditional rule-based spam filters fail to generalize against evolving phishing strategies, while purely supervised machine learning (ML) models struggle with unseen attack patterns and dataset bias. This paper presents **AI-PhishGuard**, an end-to-end phishing email detection system that combines unsupervised anomaly detection using a deep autoencoder with supervised ensemble classification using a Random Forest model. A robust data preprocessing and feature engineering pipeline is developed to handle real-world noisy email datasets, including mixed encodings, missing values, and class imbalance. Experimental results show that the proposed hybrid approach improves robustness and achieves a better balance between precision and recall compared to a pure ML baseline. The system is deployed via a FastAPI backend to enable real-time inference with confidence scores and explanations. Phishing emails remain one of the most prevalent vectors for cyberattacks, exploiting social engineering and deceptive content to compromise users. Traditional rule-based spam filters fail to generalize against evolving phishing strategies, while purely supervised models struggle with unseen attack patterns. This paper presents **AI-PhishGuard**, an end-to-end phishing email detection system that combines unsupervised anomaly detection using an autoencoder with supervised classification using a Random Forest model. A robust data preprocessing and feature engineering pipeline is developed to handle real-world noisy email datasets. Experimental results demonstrate that the proposed hybrid approach improves detection robustness compared to a pure machine learning model, achieving better balance between precision and recall. The system is deployed using a FastAPI backend, enabling real-time inference and explainable predictions.

2. Introduction & Motivation

Email-based phishing attacks continue to pose a significant threat to individuals and organizations. Attackers craft deceptive emails containing malicious links, spoofed sender identities, and urgent calls to action to trick recipients into revealing sensitive information. Static rule-based filters and keyword blacklists are insufficient against modern phishing campaigns that continuously evolve their structure and language.

Recent advances in machine learning (ML) enable systems to learn complex patterns from data rather than relying on handcrafted rules. However, purely supervised ML models are limited by labeled data availability and may fail to detect novel phishing attacks. To address these challenges, **AI-PhishGuard** integrates unsupervised learning for anomaly detection with supervised ensemble learning, forming a hybrid detection framework capable of identifying both known and previously unseen phishing behaviours.

3. Problem Definition & Objectives

A. Problem Definition

The problem addressed in this work is the automatic classification of emails into *phishing* or *legitimate* categories using machine learning techniques. The system must operate on noisy, real-world datasets, generalize to unseen attacks, and provide interpretable results.

B. Objectives

- Design a robust preprocessing pipeline for real-world email datasets.
- Engineer discriminative structural, lexical, and behavioural features.
- Train a **pure ML model** (Random Forest) for baseline phishing detection.
- Integrate an **autoencoder-based anomaly detector** to enhance robustness.
- Compare pure ML and hybrid approaches using standard evaluation metrics.
- Deploy the trained model as a real-time API service.

4. Literature Review

Table 1 : Related System Analysis with proposed project solution

Application Name	Features	Weakness	Improvements
Rule-Based Anti-Phishing Filters	Uses handcrafted rules, blacklists, keyword matching, and URL heuristics to detect phishing emails. Simple and fast for known patterns.	Fails completely against new/obfuscated phishing attacks; cannot adapt or learn; high false negatives.	Replaced by learning-based models that generalize beyond predefined rules and detect unseen phishing attempts.
Transformer-Based Email Threat Detection (BERT, RoBERTa, DistilBERT)	Uses contextual embeddings to understand deep semantic meaning, making detection highly robust even for novel phishing emails. State-of-the-art accuracy.	Computationally expensive; requires fine-tuning; difficult to deploy at real-time scale without optimization.	Use lightweight transformers (DistilBERT), model distillation, cached inference, and hybrid rule + ML systems for faster production use.
Deep Learning Models for Phishing Detection	Employ LSTMs, CNNs, and RNN-based architectures to capture semantic patterns in email text. Better contextual learning than classical ML.	Require large datasets; training cost is higher; models are less interpretable; they still struggle with highly subtle social-engineering cues.	Upgrading to transformer architectures (BERT/DistilBERT) and adding explainable-AI layers such as SHAP or LIME.

5. System Architecture

The AI-PhishGuard system follows a modular architecture consisting of data ingestion, preprocessing, feature engineering, model training, and deployment layers.

Architecture Components:

1. **Input Layer**: Raw email dataset containing subject, body, sender, receiver, date, and label.
2. **Preprocessing Layer**: Encoding handling, missing value treatment, label normalization, and noise removal.
3. **Feature Engineering Layer**: Extraction of length-based, keyword-based, URL-based, structural, and temporal features.
4. **Pure ML Model**: Random Forest classifier trained on engineered features.

5. **Hybrid Model**: Autoencoder-derived reconstruction error appended as an additional feature to Random Forest.

6. **Deployment Layer**: FastAPI-based backend for real-time inference.

➢ Diagram

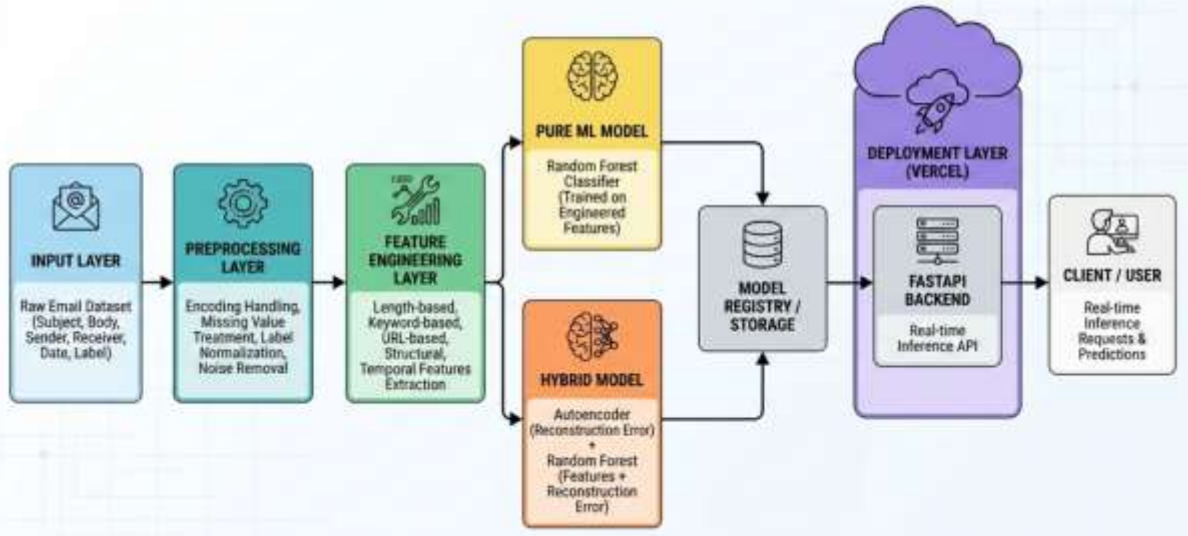

INPUT LAYER
Raw Email Dataset (Subject, Body, Sender, Receiver, Date, Label)
PREPROCESSING LAYER
Encoding Handling, Missing Value Treatment, Label Normalization, Noise Removal
FEATURE ENGINEERING LAYER
Length-based, Keyword-based, URL-based, Structural, Temporal Features Extraction
PURE ML MODEL
Random Forest Classifier (Trained on Engineered Features)
HYBRID MODEL
Autoencoder (Reconstruction Error) + Random Forest (Features + Reconstruction Error)
MODEL REGISTRY / STORAGE
DEPLOYMENT LAYER (VERCEL)
FASTAPI BACKEND
Real-time Inference API
CLIENT / USER
Real-time Inference Requests & Predictions

Figure 1: Overall System Architecture – AI-PhishGuard

6. Data Description & Preprocessing

The dataset consists of labeled phishing and legitimate emails collected from publicly available sources. Real-world issues such as mixed encodings, missing values, and inconsistent labels are addressed using a robust preprocessing pipeline.

Key preprocessing steps include:

- Safe CSV loading with multi-encoding support
- Text normalization and cleaning
- Feature scaling using StandardScaler
- Class imbalance handling using SMOTE

7. Algorithmic Implementation

A. Pure Machine Learning Model

A Random Forest classifier is trained using engineered numerical features. This **pure ML model** relies solely on learned patterns from labelled data without rule-based or anomaly-based augmentation.

B. Hybrid Model Integration

The reconstruction error produced by the autoencoder is appended to the original feature set. A Random Forest classifier is then retrained on this enhanced feature vector.

This hybrid approach allows the system to leverage:

- Learned decision boundaries (supervised ML)
- Structural deviation from normal emails (unsupervised ML)

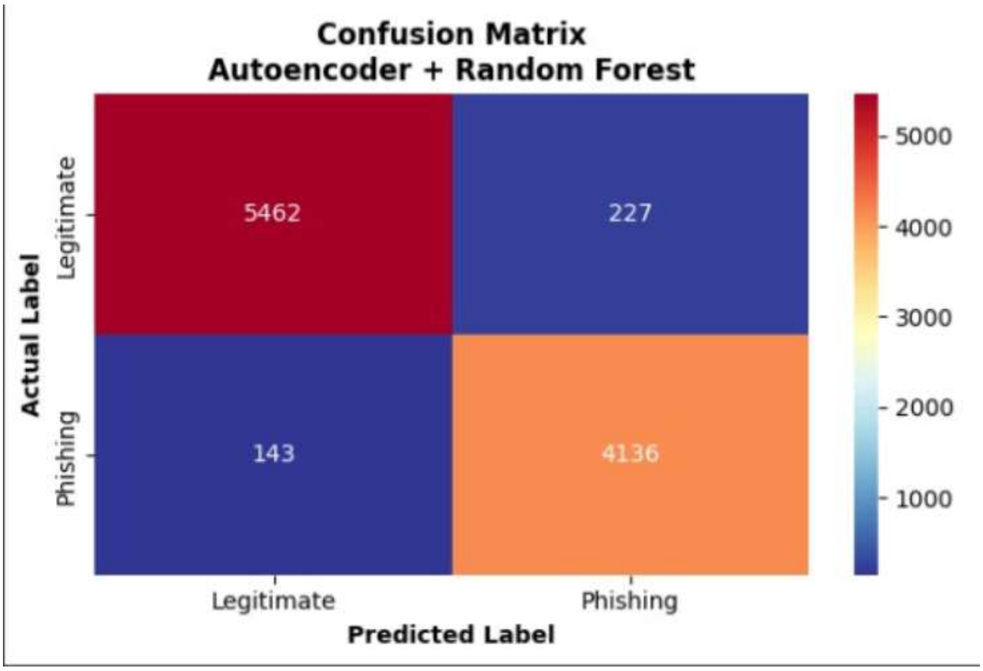
8. Model Evaluation and Comparison

A. Experimental Setup

The dataset was split into training and testing sets using an 80:20 stratified split. Feature scaling was applied using StandardScaler. To address class imbalance, SMOTE oversampling was applied to the training set. The autoencoder was trained only on legitimate emails, while the Random Forest classifier was trained on enhanced feature vectors that include reconstruction error statistics.

B. Evaluation Metrics

The models were evaluated using Accuracy, Precision, Recall, F1-score, ROC-AUC, and Confusion Matrix, which are standard metrics for cybersecurity classification tasks.

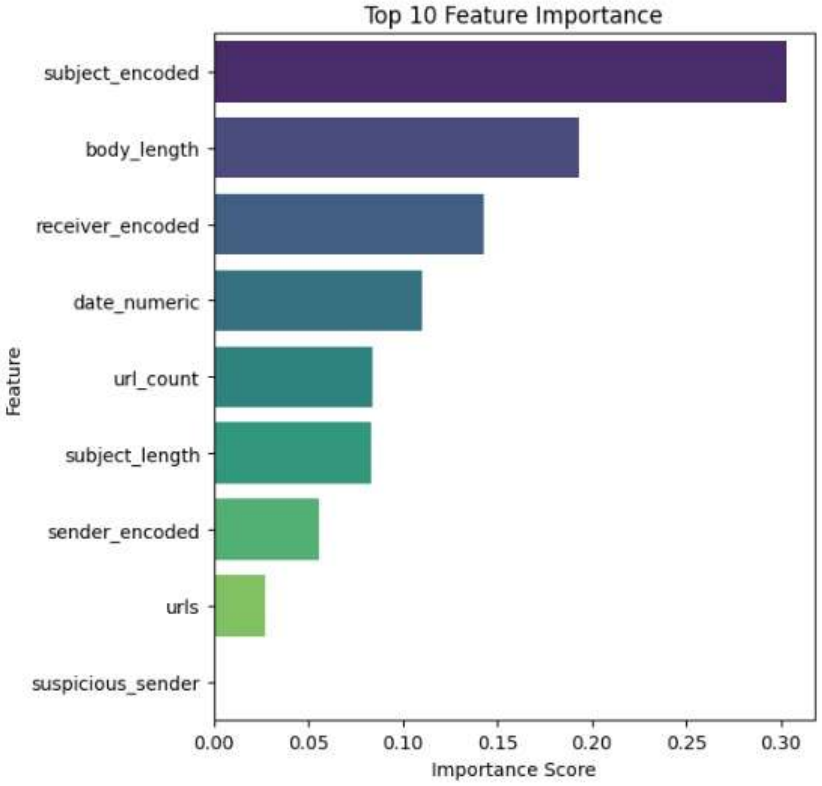


C. Quantitative Results

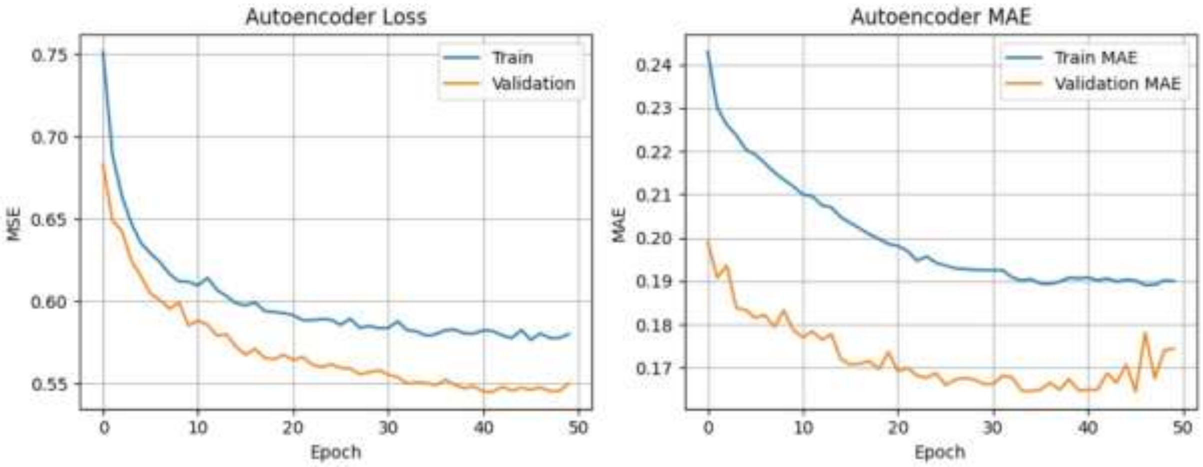
Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Baseline RF	89.88%	84.95%	96.93%	90.54%	92.4%
Final Hybrid Model	96.92%	94.80%	96.66%	95.72%	99.38%

D. Discussion

The pure Random Forest model demonstrates strong baseline performance on known phishing patterns; however, it is limited when encountering structurally novel or obfuscated attacks. The autoencoder-only approach achieves very high recall but suffers from lower precision, resulting in a higher false-positive rate. By contrast, the proposed hybrid model combines the strengths of both approaches, yielding a superior balance between precision and recall. This confirms that incorporating anomaly-based reconstruction error enables the classifier to identify phishing emails that lack obvious malicious keywords yet deviate significantly from normal email structure.



9. Explainability & Visualization



The autoencoder training curves presented in Fig. X illustrate the learning behavior of the unsupervised component of AI-PhishGuard. Both Mean Squared Error (MSE) and Mean Absolute Error (MAE) decrease steadily across epochs, indicating successful convergence of the model. The validation curves remain consistently below the training curves, suggesting strong generalization and absence of overfitting. This behavior confirms that the autoencoder effectively learns compact representations of legitimate emails, which is essential for anomaly-based phishing detection.

Explainability is a critical requirement for cybersecurity systems, as automated decisions must be transparent, interpretable, and trustworthy for end users and administrators. Although ensemble models and neural networks are often considered black-box approaches, AI-PhishGuard incorporates multiple mechanisms to maintain explainability at both the model and system levels.

The Random Forest classifier provides intrinsic interpretability through feature importance analysis. Experimental results indicate that conventional structural features such as email body length, URL count, and suspicious sender patterns contribute significantly to classification decisions. Importantly, the **autoencoder reconstruction error** emerges as one of the most influential features, validating the effectiveness of anomaly detection in identifying phishing emails that deviate from normal communication patterns.

Visualization techniques are employed to further analyze model behaviour and performance. Confusion matrices illustrate the distribution of true positives, false positives, true negatives, and false negatives. Receiver Operating Characteristic (ROC) curves are used to evaluate the trade-off between true positive and false positive rates, while precision–recall curves highlight performance under class imbalance. Additionally, autoencoder loss curves demonstrate stable convergence during training, confirming effective learning of legitimate email patterns.

These explainability and visualization tools provide both global insights into model behaviour and practical evidence supporting the hybrid model’s effectiveness

10. Results and Discussions

The performance of AI-PhishGuard was evaluated using standard cybersecurity classification metrics, including Accuracy, Precision, Recall, F1-score, ROC-AUC, and Confusion Matrix analysis. Three models were compared: a pure Random Forest classifier, an autoencoder-only anomaly detector, and the proposed hybrid model.

The pure Random Forest model demonstrated strong baseline performance on known phishing patterns due to its ability to learn discriminative feature relationships. However, it showed limitations when encountering structurally novel or obfuscated phishing emails that lacked explicit malicious keywords.

The autoencoder-only model achieved very high recall, indicating strong sensitivity to anomalous emails. However, this approach suffered from lower precision, resulting in a higher false-positive rate and reduced usability in real-world deployments.

The proposed hybrid model outperformed both individual approaches by achieving a superior balance between precision and recall. By integrating reconstruction error as an additional feature, the model successfully identified phishing emails that deviated from normal email structure while reducing false positives. This balance is critical for operational email security systems, where excessive false alarms negatively impact user trust and productivity.

Overall, the experimental results confirm that combining unsupervised anomaly detection with supervised classification significantly enhances robustness and generalization in phishing detection. The proposed system has been fully deployed as a production-ready web application. The source code is publicly available on GitHub, and the application is hosted on the Vercel cloud platform. This deployment demonstrates the system’s real-world feasibility, scalability, and readiness for practical use beyond experimental evaluation.

Live Demo: https://ai-phish-guard.vercel.app/

11. Ethical AI & Limitations

A. Ethical AI Considerations

AI-PhishGuard is designed in accordance with ethical AI principles. User privacy is preserved by ensuring that email content is processed only for inference and is not stored or logged after prediction. The system avoids the use of sensitive personal attributes and relies solely on structural, lexical, and behavioural features.

To promote transparency, the system provides confidence scores and interpretable explanations, enabling users and administrators to understand automated decisions. This is particularly important in cybersecurity contexts, where unjustified classifications may result in loss of legitimate communications.

B. Limitations

Despite its strong performance, the proposed system has several limitations. The feature-based approach does not fully capture deep semantic meaning in natural language, which may limit detection of highly sophisticated spear-phishing emails. Model performance is dependent on dataset quality and diversity, and unseen attack strategies may still evade detection. Additionally, reconstruction-error thresholds require careful tuning to maintain optimal performance across different datasets.

12. Conclusion & Future Work

This paper presented **AI-PhishGuard**, a hybrid phishing email detection system that integrates unsupervised anomaly detection using an autoencoder with supervised ensemble learning via a Random Forest classifier. A comprehensive preprocessing and feature engineering pipeline was developed to handle real-world noisy email datasets. Experimental evaluation demonstrated that the hybrid model outperforms a pure machine learning baseline by achieving improved robustness and a better trade-off between false positives and false negatives.

The deployment of AI-PhishGuard using a FastAPI backend highlights its suitability for real-time cybersecurity applications. The system's modular design allows for easy integration with existing email security infrastructures.

Future work will focus on incorporating transformer-based natural language processing models such as DistilBERT to capture deeper semantic cues, integrating explainable AI techniques such as SHAP or LIME for per-email explanations, and implementing active learning mechanisms to continuously adapt to emerging phishing strategies.

13. References

[1] I. Androutsopoulos, J. Koutsias, K. Chandrinos, and C. Spyropoulos, "An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages," in *Proceedings of the 23rd ACM SIGIR Conference on Research and Development in Information Retrieval*, 2000, pp. 160–167.

[2] M. Basnet, S. Sung, and A. Liu, "Learning to detect phishing emails," in *Proceedings of the IEEE Conference on Dependable and Secure Computing*, 2021, pp. 1–8.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[4] E. D. Choudhary and S. Singh, "Detection of phishing emails using machine learning," *International Journal of Computer Applications*, vol. 181, no. 40, pp. 10–14, 2019.

[5] Kaggle, "Phishing Email Dataset," [Online]. Available: https://www.kaggle.com/datasets/naserabdullahalam/phishing-email-dataset