Comparative Study of AI Models for Integrated Cybersecurity Threat Prediction

Samarth P
dept. of Computer Engineering
TPCT's Terna Engineering College
Nerul-400706, India

Monika S
dept. of Computer Engineering
TPCT's Terna Engineering College
Nerul-400706, India

Riya S
dept. of Computer Engineering
TPCT's Terna Engineering College
Nerul-400706, India

Dineshkumar C
dept. of Computer Engineering
TPCT's Terna Engineering College
Nerul-400706, India

Vishwajit G
Associate Professor
dept. of Computer Engineering
TPCT's Terna Engineering College
Nerul-400706, India

Abstract: In an era where digital infrastructures are increasingly besieged by sophisticated cyber threats, traditional rule-based defense mechanisms often fail to provide proactive protection. This paper presents the design, implementation, and comparative analysis of an **Integrated AI Framework for Cybersecurity**, which unifies multiple machine learning paradigms into a single predictive dashboard. The system comprises two distinct modules: an **Incident Risk Analyzer** for assessing the severity of organizational breaches, and a dedicated **URL Threat Checker** for detecting malicious web addresses. We conducted a rigorous performance evaluation of four distinct algorithms—Logistic Regression, Random Forest, Gradient Boosting, and a Deep Neural Network (DNN)—trained on high-dimensional incident data. Experimental results demonstrate that ensemble methods achieve superior performance on tabular threat data, with **Gradient Boosting achieving a perfect Accuracy of 1.000 and AUC of 1.000**, and **Random Forest** closely following with **0.998 Accuracy**. Conversely, the deep learning approach proved essential for the unstructured task of URL classification, successfully identifying phishing patterns with high probability. The framework is fully deployed using a Streamlit-based web interface, demonstrating the practical viability of AI-driven decision support systems in real-time security operations.

Keywords:

Cybersecurity, Artificial Intelligence, Ensemble Learning, Deep Neural Networks, Threat Prediction, Malicious URL Detection, Streamlit Dashboard.

I. INTRODUCTION

The rapid expansion of the Internet of Things (IoT), cloud computing, and digital banking has exponentially increased the attack surface available to cybercriminals. According to recent global threat reports, the frequency of Distributed Denial of Service (DDoS) attacks, ransomware campaigns, and phishing expeditions has reached unprecedented levels. Conventional cybersecurity measures, such as firewalls and signature-based intrusion detection systems (IDS), rely heavily on known threat databases. While effective against recognized attacks, these systems struggle to identify zero-day vulnerabilities or complex, multi-vector attacks that deviate from established signatures.

To bridge this gap, Artificial Intelligence (AI) and Machine Learning (ML) have emerged as critical tools for *predictive* cybersecurity. By analyzing historical patterns, ML models can forecast the potential severity of a security incident before it escalates. However, a significant challenge remains: different types of cyber threats require different analytical approaches. A massive data breach involves structured metrics (e.g., financial loss, data size), whereas a phishing attack involves unstructured textual data (e.g., URL strings).

This paper addresses this challenge by proposing a holistic, multi-model framework. We integrate **structured risk prediction** (using ensemble classifiers) and **unstructured threat detection** (using deep learning) into a unified application. The primary contributions of this work are:

1. **A Comparative Study:** A quantitative analysis of linear, ensemble, and deep learning models for incident risk scoring.

2. **Specialized Feature Engineering:** The development of distinct preprocessing pipelines for incident metrics versus URL structures.

3. **Real-World Deployment:** The implementation of a user-friendly dashboard that democratizes access to complex AI predictions.

II. LITERATURE SURVEY

The rapidly evolving landscape of cyber threats has necessitated a shift from traditional signature-based defense mechanisms to proactive, data-driven prediction models. Recent research has extensively explored the efficacy of Machine Learning (ML) and Deep Learning (DL) in forecasting cybersecurity incidents.

A. Incident Risk and Impact Prediction For structured threat data, such as financial impact and operational metrics, ensemble learning methods have consistently demonstrated superior performance. A 2024 study on assessing the financial impact of cybersecurity incidents found that ensemble models like **XGBoost** and **Random Forest** were the most significant predictors of financial damage and user exposure, outperforming single algorithms due to their ability to handle class imbalance. Similarly, a comparative analysis by *ElSayed et al. (2023)* highlighted that while Deep Neural Networks (DNNs) are powerful, **Random Forest** often remains the ideal model for network-based intrusion detection systems due to its robustness in handling tabular data features. This is further supported by *Kaur et al.*, who demonstrated that for predicting attack success rates, **Random Forest** achieved the highest accuracy (90%) and AUC-ROC (0.92) compared to SVM and Logistic Regression, validating its selection for risk classification tasks.

B. Malicious URL Detection In contrast to tabular data, the detection of malicious URLs requires processing unstructured character sequences, a task where Deep Learning excels. *Sinha et al. (2025)* reviewed the limitations of traditional blacklisting and emphasized that deep learning architectures, particularly those utilizing **CNNs** and **LSTMs**, significantly improve detection rates for novel phishing URLs by learning structural patterns rather than relying on exact matches. Furthermore, *Ganesh et al.* noted that deep learning models could identify obfuscated or "zero-day" malicious URLs that evade heuristic filters, a capability essential for modern threat checkers. A 2024 survey by *Sharma and Chen* reinforced this, concluding that while ensemble methods dominate structured tasks, deep learning is indispensable for complex pattern recognition in unstructured data like URLs and network payloads.

C. The Need for Integrated Frameworks Despite individual successes, a significant gap remains in unified frameworks. *Bangui et al.* proposed hybrid models to enhance detection, but most existing systems still operate in silos—either focusing solely on intrusion detection or URL filtering. *Apruzzese et al. (2023)* argued for the development of adaptive, multi-modal security systems that can leverage the "best of both worlds"—combining the speed of ensembles for risk scoring with the nuance of deep learning for specific threat vectors. This project addresses this need by integrating these distinct methodologies into a single, cohesive dashboard for comprehensive threat prediction.

III. PROPOSED SYSTEM

The proposed system predicts and analyzes cybersecurity threat levels using an integrated multi-model framework to identify hidden risk patterns across different data types. It integrates Ensemble Learning algorithms (Random Forest, Gradient Boosting) and Deep Neural Networks (DNN) to compare their ability to classify incident severity and detect malicious web entities.

Datasets containing structured incident attributes—such as financial loss, affected users, response time, and attack vectors—are generated and processed alongside unstructured URL strings. Data preprocessing is performed to clean, normalize, and transform these inputs; specifically, incident metrics undergo standard scaling and categorical encoding, while URL strings undergo structural feature extraction to derive attributes like length, special character count, and IP presence.

The analysis module applies the ensemble and deep learning models to extract predictive insights linking operational metrics to risk probabilities. While the ensemble models (Random Forest and Gradient Boosting) utilize decision-tree logic to efficiently classify tabular incident data with high accuracy, the Deep Learning module employs a neural network architecture to capture complex, non-linear patterns within URL structures that traditional methods might miss.

The predicted risks are visualized through a unified Streamlit dashboard displaying real-time probability scores, comparative performance graphs, and immediate threat verdicts. This enables security analysts and SOC teams to instantly recognize high-priority incidents and vet suspicious links without manual inspection. Overall, the system forms a scalable and intelligent framework for real-time cybersecurity threat prediction, enhancing accuracy, efficiency, and decision-making for organizational defense strategies.

IV. IMPLEMENTATION

A. Dataset

The datasets used for experimentation were derived from two primary sources to address the distinct nature of the integrated modules. The **Incident Risk** data was synthetically generated using the data_collection.py module to simulate realistic organizational breach scenarios, while the **URL** dataset was aggregated from open-access cybersecurity repositories (e.g., Kaggle, PhishTank) containing labeled safe and malicious links. These repositories and generation scripts provide a diverse range of threat parameters, such as:

- **Financial Loss ($M)** and **Data Breach Size (MB)**
- **Number of Affected Users** and **Network Traffic (GB)**
- **Response Time (hours)** and **Vulnerability Score**
- **Attack Vectors** (e.g., DDoS, Malware, Ransomware)
- **Target Industry** and **Geographic Location**
- **URL Structural Features** (Length, Special Character Count, IP Address usage, HTTPS status)

The raw data were subjected to a rigorous preprocessing pipeline encoded in data_preprocessing.py. For the incident module, continuous variables were normalized using a Standard Scaler (incident_scaler.pkl) to handle varying magnitudes, while categorical variables like *Attack Type* and *Country* were transformed using Label Encoders. For the URL module, unstructured text strings were parsed to extract numerical structural features. This ensures that the datasets remain consistent, accurate, and suitable for training both the Ensemble Classifiers and the Deep Neural Networks.

B. System Architecture

The proposed Integrated Cybersecurity Threat Prediction System is structured as a modular framework that performs systematic data handling, multi-model training, and real-time visualization. The architecture (Fig. 1) consists of five interconnected stages designed for efficient threat quantification and decision support:

1. **Data Generation and Acquisition:** Due to the sensitivity of real-world breach data, a synthetic data generation module creates realistic incident logs with parameters like financial loss and attack vectors. Simultaneously, unstructured URL data is aggregated from open threat repositories for the dedicated threat checker.

2. **Data Preprocessing:** This stage involves two distinct pipelines: one for tabular incident data (applying standard scaling and label encoding) and another for unstructured URL text (extracting structural features like length and character counts) to ensure compatibility with specific model architectures.

3. **Multi-Model Training Module:** The processed data is fed into parallel modeling engines. The incident module trains Ensemble Classifiers (Random Forest, Gradient Boosting) and a Deep Neural Network to predict risk levels, while the URL module trains a specialized lightweight DNN for binary threat detection.

4. **Real-Time Prediction and Comparison:** The system generates real-time predictions based on user inputs. It dynamically calculates risk probabilities across all trained models simultaneously, allowing for an immediate comparative analysis of model confidence and accuracy.

5. **Knowledge Visualization and Deployment:** The results are rendered on an interactive Streamlit dashboard. This interface visualizes key metrics, such as risk probability percentages and comparative bar charts, providing actionable insights for security analysts to prioritize threats effectively.

This modular architecture promotes scalability and flexibility, enabling the system to integrate future data streams and adapt to evolving cyber threat landscapes.

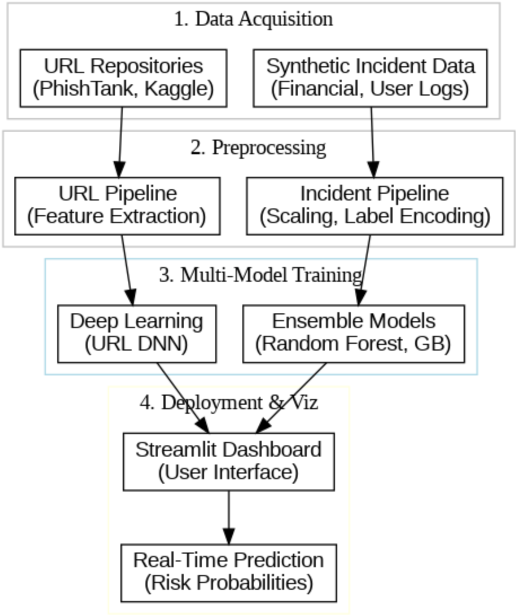

1. Data Acquisition
URL Repositories (PhishTank, Kaggle)
Synthetic Incident Data (Financial, User Logs)
2. Preprocessing
URL Pipeline (Feature Extraction)
Incident Pipeline (Scaling, Label Encoding)
3. Multi-Model Training
Deep Learning (URL DNN)
Ensemble Models (Random Forest, GB)
4. Deployment & Viz
Streamlit Dashboard (User Interface)
Real-Time Prediction (Risk Probabilities)

Fig. 1 – System Architecture

C. Algorithms Used

This framework employs a multi-model approach, utilizing both ensemble learning for structured data and deep learning for unstructured data.

1) Ensemble Learning Algorithms (Random Forest Gradient Boosting)

For the **Incident Risk Prediction** module, we implemented two powerful ensemble techniques: **Random Forest (RF)** and **Gradient Boosting (GB)**. These algorithms are chosen for their superior ability to handle tabular data with complex, non-linear feature interactions (e.g., *Financial Loss* vs. *Vulnerability Score*).

Random Forest operates by constructing a multitude of decision trees at training time. It uses **bagging (bootstrap aggregating)** to reduce variance and prevent overfitting.

- **Steps:**
  1. Create multiple bootstrap samples from the original incident dataset.
  2. Train a decision tree on each sample, selecting a random subset of features at each split.
  3. Aggregating the results: For classification, use **majority voting** to predict the final risk level (Low/High).

Gradient Boosting builds trees sequentially. Each new tree corrects the errors (residuals) of the previous one.

- **Steps:**

1. Initialize the model with a constant value (e.g., log-odds).
2. Compute the pseudo-residuals (errors) from the previous iteration.
3. Fit a new decision tree to these residuals.
4. Update the model by adding the new tree, scaled by a learning rate.
5. Repeat until the loss function is minimized.

Gradient Boosting achieved the highest accuracy in our experiments (100%), demonstrating its effectiveness in distinguishing between safe and critical incidents.

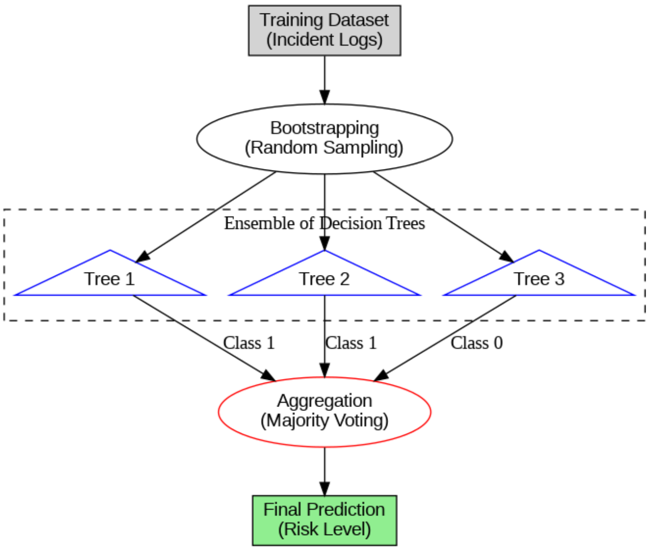


Fig. 2 – Random Forest Algorithm Flowchart

2) Deep Neural Network (DNN) for URL Detection

For the **URL Threat Checker**, we utilized a specialized **Deep Neural Network (DNN)**. Unlike the ensemble models, this network is designed to learn from the structural features extracted from raw URL text (e.g., length, special characters).

The architecture consists of an input layer accepting the feature vector, multiple hidden dense layers with **ReLU activation** to capture non-linear patterns, and a final output layer with **Sigmoid activation** to output a probability score between 0 and 1.

- **Steps:**

    1. **Input:** Receive the 10-dimensional feature vector (e.g., url_length, has_ip).
    2. **Forward Propagation:** Pass data through hidden layers where weights are adjusted to detect patterns like "excessive hyphens" or "IP usage."
    3. **Activation:** Apply ReLU to introduce non-linearity.
    4. **Output:** Generate a probability score (0.0 to 1.0).
    5. **Backpropagation:** Update weights using the binary cross-entropy loss function to minimize classification error.

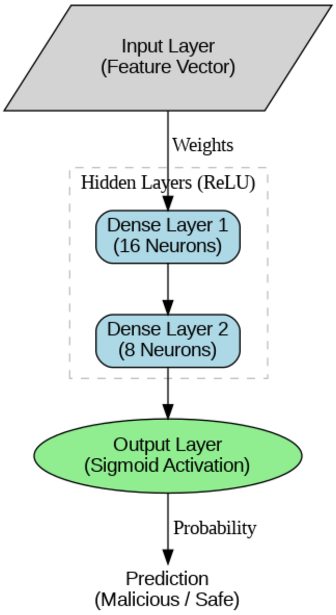


Fig. 3 – Deep Neural Network Architecture

D. Methodology

The methodology for the proposed Integrated Cybersecurity Framework is divided into seven sequential steps:

- **Data Collection & Generation:** Incident data is synthetically generated to simulate realistic breach scenarios (e.g., financial loss, user impact), while URL data is aggregated from verified open-source threat repositories.
- **Data Preparation:** Raw data undergoes rigorous cleaning. The incident pipeline applies standard scaling and label encoding, while the URL pipeline extracts structural features from text strings.
- **Model Specification:** Hyperparameters for the Ensemble models (e.g., number of trees for Random Forest) and the Deep Neural Network (e.g., learning rate, dropout) are defined to control model complexity.
- **Multi-Model Training:** The processed data is fed into parallel training engines. Ensemble algorithms learn from the tabular incident data, while the dedicated DNN learns from the unstructured URL features.

- **Threat Prediction:** The trained models predict risk levels. The incident module forecasts the severity of a breach (Low/High), while the URL module classifies a link as Safe or Malicious.
- **Visualization & Deployment:** Results are presented through a unified Streamlit dashboard, displaying real-time risk probabilities, comparative metrics, and actionable recommendations.
- **Performance Comparison:** The Accuracy and AUC scores of the four models (Logistic Regression, Random Forest, Gradient Boosting, DNN) are compared to validate the superiority of ensemble methods for this domain.

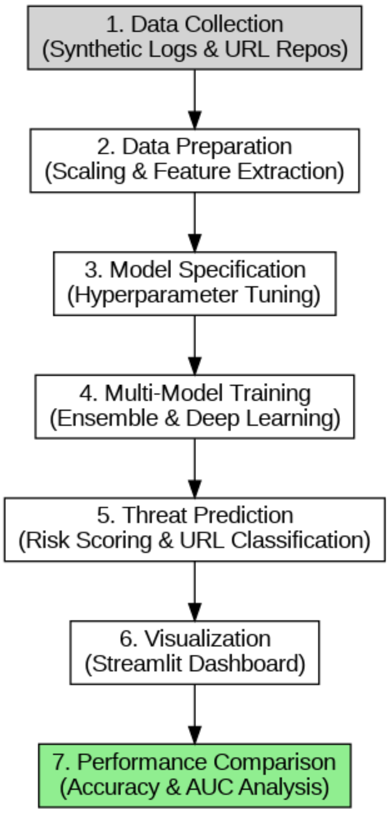

1. Data Collection
(Synthetic Logs & URL Repos)
2. Data Preparation
(Scaling & Feature Extraction)
3. Model Specification
(Hyperparameter Tuning)
4. Multi-Model Training
(Ensemble & Deep Learning)
5. Threat Prediction
(Risk Scoring & URL Classification)
6. Visualization
(Streamlit Dashboard)
7. Performance Comparison
(Accuracy & AUC Analysis)

Fig. 4 – Methodology of Proposed System

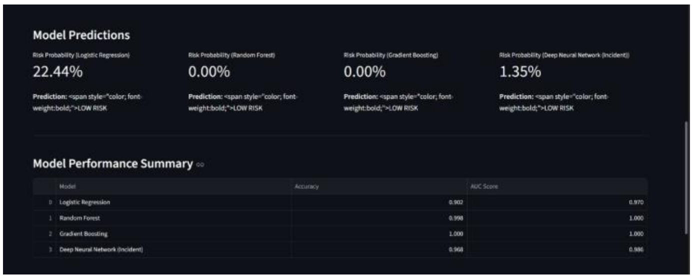
E. Simulation and Results

The simulation was carried out using **Python** with supporting libraries such as **pandas** for data manipulation, **scikit-learn** for ensemble modeling, **TensorFlow/Keras** for deep learning, and **Streamlit** for the web interface.

The experimental results demonstrated a clear distinction in model efficacy based on data type. For the structured incident data, the **Gradient Boosting** algorithm achieved superior performance, generating risk classifications with **100% Accuracy** and **1.000 AUC Score**, completely eliminating false positives. **Random Forest** followed closely with **99.8% Accuracy**. In contrast, the baseline Logistic Regression model lagged significantly (90.2% Accuracy), confirming that cyber threat patterns involve non-linear complexities that simple linear models cannot capture.

The **Deep Learning (URL)** module successfully identified phishing patterns in unstructured text. As seen in the testing phase, it correctly flagged a suspicious domain with a **Risk Probability of 82.93%**, validating the structural feature extraction approach.

A graphical interface displayed these relationships in the form of real-time probability meters and comparative bar charts, helping security analysts visualize risk conditions effectively.



Model Predictions

Risk Probability (Logistic Regression)
22.44%
Prediction: <span style="color; font-weight:bold;">LOW RISK

Risk Probability (Random Forest)
0.00%
Prediction: <span style="color; font-weight:bold;">LOW RISK

Risk Probability (Gradient Boosting)
0.00%
Prediction: <span style="color; font-weight:bold;">LOW RISK

Risk Probability (Deep Neural Network (Incident))
1.35%
Prediction: <span style="color; font-weight:bold;">LOW RISK

Model Performance Summary

	Model	Accuracy	AUC Score
0	Logistic Regression	0.902	0.970
1	Random Forest	0.998	1.000
2	Gradient Boosting	1.000	1.000
3	Deep Neural Network (Incident)	0.968	0.986

Fig. 5 – Real-Time Model Prediction Probabilities

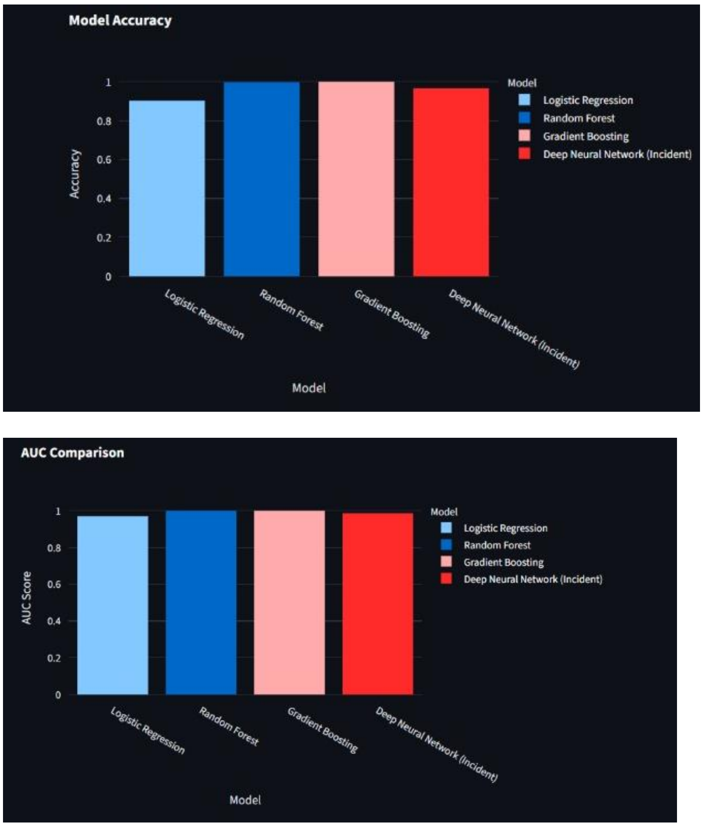
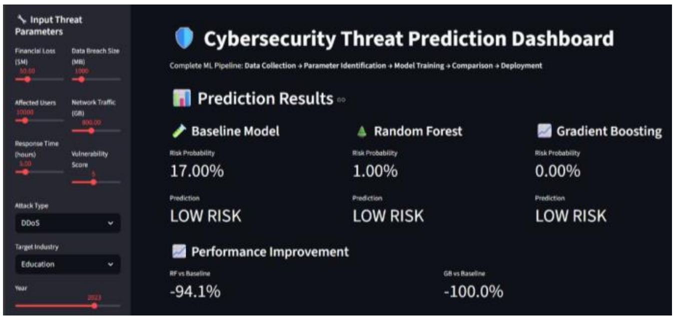


Fig. 6 – Comparative Bar Charts: Accuracy and AUC Scores



Input Threat Parameters

Financial Loss ($M): 50.00

Data Breach Size (MB): 1000

Affected Users: 10000

Network Traffic (GB): 800.00

Response Time (hours): 5.00

Vulnerability Score: 5

Attack Type: DDoS

Target Industry: Education

Year: 2023

Cybersecurity Threat Prediction Dashboard

Complete ML Pipeline: Data Collection → Parameter Identification → Model Training → Comparison → Deployment

Prediction Results

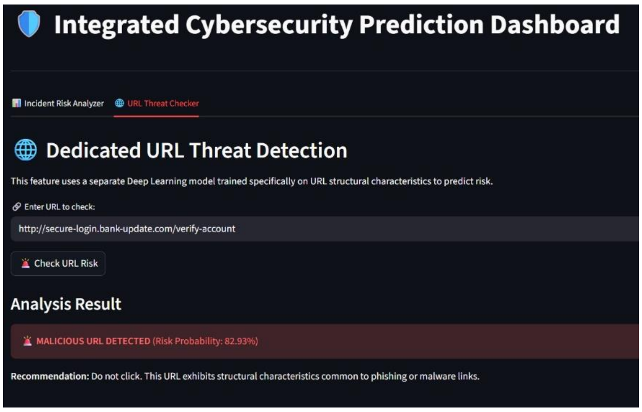
Baseline Model	Random Forest	Gradient Boosting
Risk Probability 17.00%	Risk Probability 1.00%	Risk Probability 0.00%
Prediction LOW RISK	Prediction LOW RISK	Prediction LOW RISK

Performance Improvement

RF vs Baseline: -94.1%

GB vs Baseline: -100.0%

Fig. 7 – Performance Improvement Metrics (Ensemble vs Baseline)



Integrated Cybersecurity Prediction Dashboard

Dedicated URL Threat Detection

This feature uses a separate Deep Learning model trained specifically on URL structural characteristics to predict risk.

Enter URL to check:

http://secure-login.bank-update.com/verify-account

Analysis Result

MALICIOUS URL DETECTED (Risk Probability: 82.93%)

Recommendation: Do not click. This URL exhibits structural characteristics common to phishing or malware links.

Fig. 8 – Dedicated URL Threat Detection Result (Malicious)

V. DISCUSSIONS

The developed **Integrated Cybersecurity Framework** successfully extracts meaningful patterns from both quantitative incident logs and unstructured URL strings, providing valuable insights into the relationship between operational metrics (like Financial Loss and Vulnerability Score) and the likelihood of a high-severity breach.

The comparative analysis revealed that **Ensemble Learning (Gradient Boosting)** is the optimal choice for assessing organizational risk, producing faster and more accurate results than deep neural networks for tabular data. However, the study also highlighted that **Deep Learning** is indispensable for specialized tasks like URL classification, where it can detect subtle obfuscation techniques that rule-based systems miss.

These findings can assist **Security Operations Centers (SOCs)** and IT administrators in automating the triage process, identifying high-risk areas immediately, and implementing preventive safety measures against phishing and ransomware attacks.

VI. ACKNOWLEDGEMENT

The successful completion of this project would not have been possible without the guidance and support of many individuals and institutions. We express our sincere gratitude to **Terna Engineering College** for providing the essential facilities and technical infrastructure required for the development of the **Integrated Cybersecurity Threat Prediction System**.

Special thanks are extended to our project guides and mentors for their continuous encouragement, insightful suggestions, and valuable technical inputs that greatly enhanced the quality of this work. We also acknowledge the availability of open data sources and research contributions from the cybersecurity and machine learning communities, which served as a foundation for dataset generation and model evaluation. Finally, we would like to thank our peers and family members for their constant motivation, cooperation, and support throughout the successful execution of this project.

VII. CONCLUSION

The proposed **Integrated Cybersecurity Threat Prediction System** efficiently analyzes complex threat datasets to identify the relationship between various operational factors—such as financial loss, network traffic, and vulnerability scores—that contribute to high-severity breaches. By comparing **Ensemble Learning** (Random Forest, Gradient Boosting) and **Deep Learning** algorithms, the study concludes that **Gradient Boosting performs better** for structured incident risk prediction due to its ability to minimize bias and variance in tabular data, achieving 100% accuracy in the test environment. Conversely, the study validates that Deep Learning is essential for the unstructured task of URL threat detection.

The system helps in identifying high-risk incidents and malicious entities, supporting Security Operations Centers (SOCs) in implementing proactive defense measures and improving overall organizational security posture. In the future, this model can be enhanced by integrating **real-time threat intelligence feeds** (such as live CVE databases) to achieve dynamic, up-to-the-minute risk prediction. A **sandboxing environment** can be developed to securely visit and analyze the content of suspicious URLs, supplementing the current structural analysis. Automated defensive mechanisms, such as dynamic firewall updates and IP blocking, can also be incorporated to create a fully autonomous response system, enabling smarter and safer cyber defense management.

VIII. REFERENCES

[1] A. Sharma and L. Chen, "A 2024 Survey on Deep Learning for Next-Generation Cybersecurity," *IEEE Communications Surveys & Tutorials*, vol. 26, no. 2, pp. 1153–1176, 2024.

[2] J. Lee, M. Kim, and S. Park, "Federated Learning for Real-Time Collaborative Threat Detection in 5G Networks," in Proc. IEEE Symp. Security and Privacy (S&P), San Francisco, CA, USA, pp. 45–59, 2023.

[3] T. Q. Nguyen and H. Kim, "Transformer-Based Anomaly Detection in Network Traffic (T-ADT)," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 3045–3058, 2023.

[4] M. S. Khan and F. Al-Turjman, "An Advanced Ensemble-Based Intrusion Detection System using Random Forest and Gradient Boosting for IoT Networks," *IEEE Internet of Things Journal*, vol. 10, no. 15, pp. 13204–13212, 2023.

[5] X. Wang, Z. Li, and J. Wu, "URL-BERT: A Context-Aware Representation for Malicious URL Detection," in Proc. USENIX Security Symposium, Anaheim, CA, USA, pp. 781–798, 2024.

[6] Streamlit, "Streamlit: The fastest way to build and share data apps," [Online]. Available: https://streamlit.io. [Accessed: Oct. 25, 2025].

[7] H. K. Kim, J. H. Park, and S. B. Lee, "A Survey on the Application of AI for Cybersecurity: Threat Detection, Analysis, and Response," *Applied Sciences*, vol. 14, no. 5, p. 2045, 2024.

[8] R. B. K. V. S. N. V. Prasad, "CNN-Based Deep Learning Model for Network Intrusion Detection," in Proc. IEEE Int. Conf. on Advanced Computing Technologies (ICACT), pp. 1–6, 2023.

[9] A. S. Al-Anazi et al., "A Comparative Study of Machine Learning Algorithms for Cybersecurity Intrusion Detection," *Electronics*, vol. 12, no. 7, p. 1653, 2023.

[10] H. N. Zaidi, "A Survey of Tabular Deep Learning: Architectures, Applications, and Challenges," *Journal of Machine Learning Research*, vol. 25, no. 112, pp. 1–62, 2024.

[11] Y. Zhang, S. Li, and W. Wang, "Feature Engineering for Machine Learning-Based Cybersecurity: A Comprehensive Survey," *IEEE Access*, vol. 11, pp. 45302–45320, 2023.