A DeBERT and Greedy Tree Classifier Framework for Multivariate Cyber Threat Prediction in Healthcare Security

M. Rakesh[1*], Gangadhari Vivek[2], Kashireddy Varun Reddy[2], Alpana Venkat Ashish[1]

[1]Assistant Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering

[1,2]Kommuri Pratap Reddy Institute of Technology, Ghanpur, Ghatkesar, 501301, Telangana, India.

*Correspondence: M. Rakesh (machrak3149@gmail.com)

To Cite this Article
M. Rakesh, Gangadhari Vivek, Kashireddy Varun Reddy, Alpana Venkat Ashish, "A DeBERT and Greedy Tree Classifier Framework for Multivariate Cyber Threat Prediction in Healthcare Security", Journal of Science Engineering Technology and Management Science, Vol. 03, Issue 04, April 2026,pp: 898-910, DOI: http://doi.org/10.64771/jsetms.2026.v03.i04.pp898-910
Submitted: 08-03-2026 Accepted: 16-04-2026 Published: 23-04-2026

ABSTRACT

The rapid digital transformation of healthcare, driven by the adoption of Electronic Health Records (EHRs), interconnected medical devices, and integrated information systems, has significantly enhanced operational efficiency and patient care. However, this increased interconnectivity has also expanded the attack surface, exposing healthcare infrastructures to a wide range of cybersecurity threats. Factors such as legacy system dependencies, misconfigurations, and delayed software updates further intensify the risk of data breaches and service disruptions. Traditional manual methods for analyzing security logs and incident reports are increasingly inadequate due to the growing volume and unstructured nature of data, leading to delayed and error-prone threat detection. To address these challenges, this work introduces an automated framework based on Natural Language Processing (NLP) for efficient cyber threat identification and vulnerability analysis. The approach begins with robust text preprocessing, including normalization, tokenization, lemmatization, and removal of irrelevant terms to enhance data quality. Contextual feature extraction is performed using the transformer-based RoBERTa model, enabling deeper semantic interpretation of cybersecurity text. To mitigate class imbalance and improve model robustness, the Adaptive Synthetic Sampling (AdaSYN) technique is applied. The system evaluates multiple machine learning classifiers, including Greedy Tree Classifier (GTC), Tao Tree Classifier (TTC), K-Nearest Neighbors (KNN), and Gaussian Naive Bayes (GNB), facilitating comprehensive performance comparison. Experimental results identify the optimized GTC model as the most effective in classifying threat types, assessing severity, and recommending mitigation strategies. The solution is implemented as a secure web-based application that supports streamlined data input, automated analysis, and intuitive result visualization. The proposed framework enhances detection accuracy, ensures interpretability, and enables proactive cybersecurity management, thereby strengthening the resilience of modern healthcare systems.

Keywords: Electronic health records, Cybersecurity threats, Vulnerability Assessment, Real-Time Threat Analysis, Natural Language Processing (NLP).

1. INTRODUCTION

The rapid digitization of healthcare systems has brought transformative advantages, including greater efficiency in clinical operations, reduced costs, improved patient safety, and enhanced overall quality of care. While these advancements have significantly strengthened healthcare delivery, they have also introduced complex cybersecurity challenges that cannot be overlooked. The integration of digital technologies and interconnected systems has expanded the attack surface within Healthcare Information Infrastructure (HCII), creating new opportunities for cyber adversaries to exploit system vulnerabilities. In recent years, the healthcare sector has become an increasingly attractive target for cyberattacks, with a large proportion of organizations reporting data breaches, highlighting the growing severity of the issue. The proliferation of interconnected medical devices further intensifies these risks, as a single compromised device can potentially affect the entire network ecosystem.

Real-world incidents as shown in fig 1 underscore the seriousness of these vulnerabilities. Security flaws discovered in devices such as infusion pumps and insulin delivery systems have revealed the potential for direct threats to patient safety, while experimental attacks on critical devices like pacemakers and implantable cardiac defibrillators demonstrate the potentially life-threatening consequences of cyber intrusions. The emergence of Medical Internet of Things (IoT) devices has further complicated the cybersecurity landscape, as these devices often lack robust security mechanisms, making them prime targets for exploitation. Additionally, human-related factors, such as lack of awareness, poor security practices, and insider risks, play a significant role in compromising system security. Consequently, gaining a comprehensive understanding of threats and vulnerabilities within healthcare systems is essential for implementing effective security controls and ensuring system resilience.

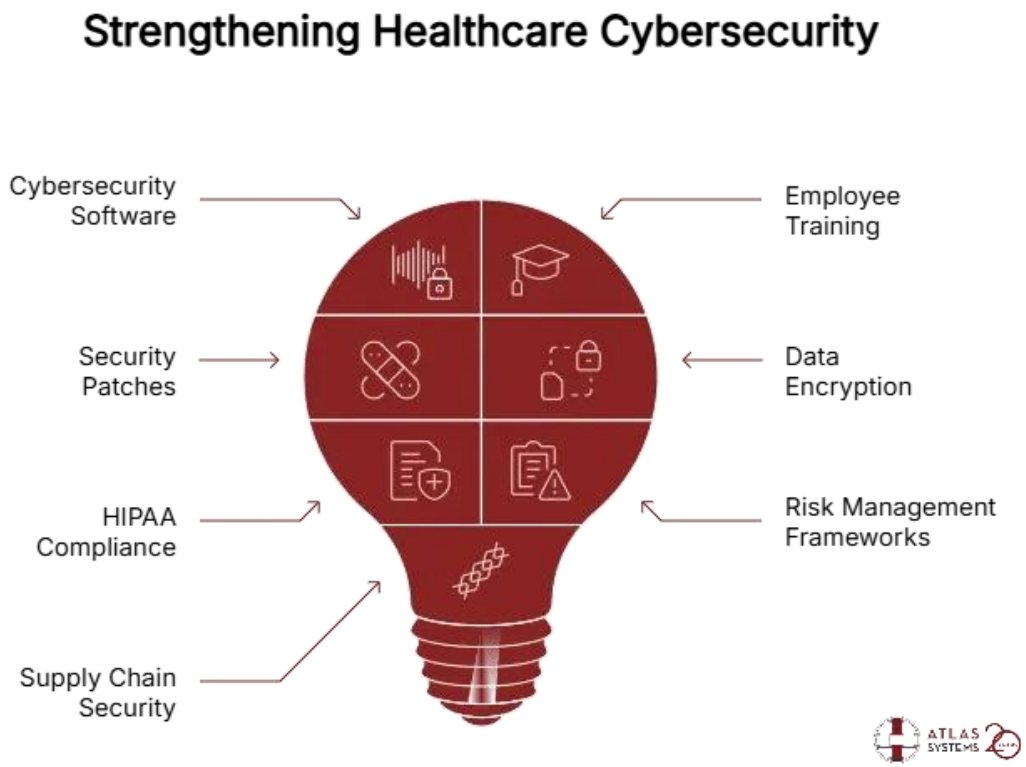

Fig 1. Cyber security threats in healthcare

Despite the urgency, analyzing cybersecurity threats in healthcare remains a complex and resource-intensive task. The vast number of reported vulnerabilities and the challenge of identifying relevant

information within a healthcare-specific context make manual analysis inefficient and error-prone. A considerable amount of valuable cybersecurity information exists in unstructured natural language formats across various online sources, including technical blogs, cybersecurity reports, news articles, and social media platforms. These sources contain critical insights into threats, attack patterns, and vulnerabilities affecting HCII assets. However, extracting actionable knowledge from such data is inherently difficult due to the complexity of natural language, which includes ambiguity, varied contextual meanings, long and intricate sentence structures, and domain-specific terminology filled with abbreviations and technical jargon. These challenges necessitate the adoption of advanced analytical approaches to effectively process and interpret cybersecurity-related textual data in the healthcare domain.

The healthcare ecosystem generates large volumes of cyber-threat data from logs, incident reports, and security alerts, but traditional systems rely on manual analysis, static rule-based detection, and fragmented monitoring tools, which are slow, error-prone, and unable to detect evolving attacks such as ransomware, phishing, malware, and DDoS. These conventional approaches lack the ability to process unstructured textual threat information, fail to classify attacks by severity or type, and provide limited support for proactive threat prioritization, resulting in delayed response, higher vulnerability exposure, and increased risk to patient data and clinical operations. To overcome these challenges, the proposed system introduces an NLP-based automated framework that preprocesses cyber-threat text, extracts semantic features using RoBERTa embeddings, and applies machine-learning models such as GTC, TTC, KNN, and GNB to classify threat category, estimate severity level, and suggest suitable defence mechanisms, thereby improving accuracy, automation, and security decision-making in healthcare environments.

2. LITERATURE SURVEY

There are several recent works that focus on threat and vulnerability detection and analysis based on Machine Learning (ML) models. In Ghaffarian et al. [1], a survey of ML and Data Mining techniques to mitigate the damages of software vulnerabilities is presented. In Satyapanich et al. [2], a semantic schema to describe CS events was presented using Deep Learning-based Information Extraction (IE) pipeline to implement the automatic extraction of structured information about data breaches, ransomware and phishing attacks and the discovery and the patches of vulnerabilities. In Gao et al. [3], a data and knowledge-driven CS Named Entity Recognition (NER) method is presented, exploiting a Bidirectional Long Short Term Memory with Conditional Random Field (BiLSTM-CRF) architecture, including also a multi-head self-attention neural network with word embeddings trained on CS closed-domain texts to improve their effectiveness in conjunction with KBs, for the recognition of the details of the assets (application, vendor, version, etc.) involved in CS issues. In Nikoloudakis et al. [4], a ML-based situational awareness framework is presented which is able to detect existing and newly introduced network-enabled entities in an IoT-based environment based on real-time awareness features provided by the Software-Defined Networking (SDN) paradigm, assessing them against known vulnerabilities, and assigning them to a connectivity-appropriate network slice. The authors of [5] developed software vulnerability detection as an NLP problem with source code treated as texts, addressing the automated software vulnerability detection using recent DL NLP models. They compared various DL models based on their accuracy and the best performer achieved 95% of accuracy. Furthermore, the proposed approach was also able to predict the vulnerability class of source codes. The authors of [6] presented an NLP DL-based architecture for the identification of relevant CS information, such as vulnerability exploitations, attack discoveries and advanced persistent threats. This architecture is composed of a word-embedding layer, a BiLSTM layer, and a CRF layer, concatenated with a further BiLSTM as output layer. The results of their experiments

showed some improvements with respect to the baselines. The authors of one paper [7] presented a method to analyze the severity of CS threats analyzing the language of CS-related tweets through a DL approach. The experiments used a corpus of 6000 tweets containing the description of software vulnerabilities, annotated with the opinions of the authors toward their severity. The paper also presented a method for linking software vulnerabilities reported in tweets to CVEs and NVD KBs. The obtained results demonstrated a high-precision in forecasting high-severity vulnerabilities, also highlighting that reports of severe vulnerabilities extracted from online sources are predictive of real-world exploits. According to [8], researchers need to deeply investigate ethical compliance even when the data seem to be public. Usually, in CS research the data are accessed and analyzed without the informed consent of participants, but acquiring informed consent could be practically impossible with datasets containing hundreds of data. In the case of the experimental assessment presented in this work, there is no personal data included, so there are no ethical issues. The authors of [9] presented a method for NER in the CS domain that uses a model that integrates BERT and BiLSTM-CRF DL architectures, improving baseline performance. As per recent study showed that at least 20% of the medical device manufacturers experienced ransomware or malware attacks in the last 20 months. The authors of [10] proposed a cyber supply chain threat analysis that integrates Random Forest and XGBoost algorithms for the threat prediction. The work considers threat intelligence and predicts the Tactics, Techniques, and Procedures (TTP) deployed for a cyber attack, demonstrating high accuracy in their experimental assessment. The authors of [8] reviewed and compared different generic cyber risk assessment frameworks in the healthcare field, comparing them, discussing the methodology of assessment and the limitations associated with them. In [13] is presented SecureBERT, a Bidirectional Encoder Representations from Transformers (BERT) model trained on CS-domain large NL corpora, which outperforms other similar models in NLP tasks in the CS domain. The authors of [10] collected a large corpus of labeled sequences from Industrial Control Systems device's documentation to pre-train and fine-tune a BERT language model, named CyBERT.

3. PROPOSED SYSTEM

The proposed system presents an advanced Machine Learning pipeline for NLP-based analysis of cyber threats and vulnerabilities within the healthcare ecosystem. The workflow begins with preprocessing raw textual threat data using NLTK for cleaning and lemmatization, while simultaneously encoding the three target variables: Threat Category, Severity Score, and Suggested Defense Mechanism. As shown as Fig 2. The system leverages a Transformer-based feature extraction approach using RoBERTa to generate dense, context-aware embeddings that effectively capture semantic relationships within the text, overcoming the limitations of traditional methods. To address class imbalance, the dataset is balanced using SMOTE-based resampling techniques. The extracted features are then utilized by multiple supervised classifiers, including KNN, GNB, TTC, and the proposed GTC. Among these, GTC is selected as the final model due to its strong predictive performance and interpretability, enabling accurate, fast, and explainable threat classification for actionable cybersecurity decisions.

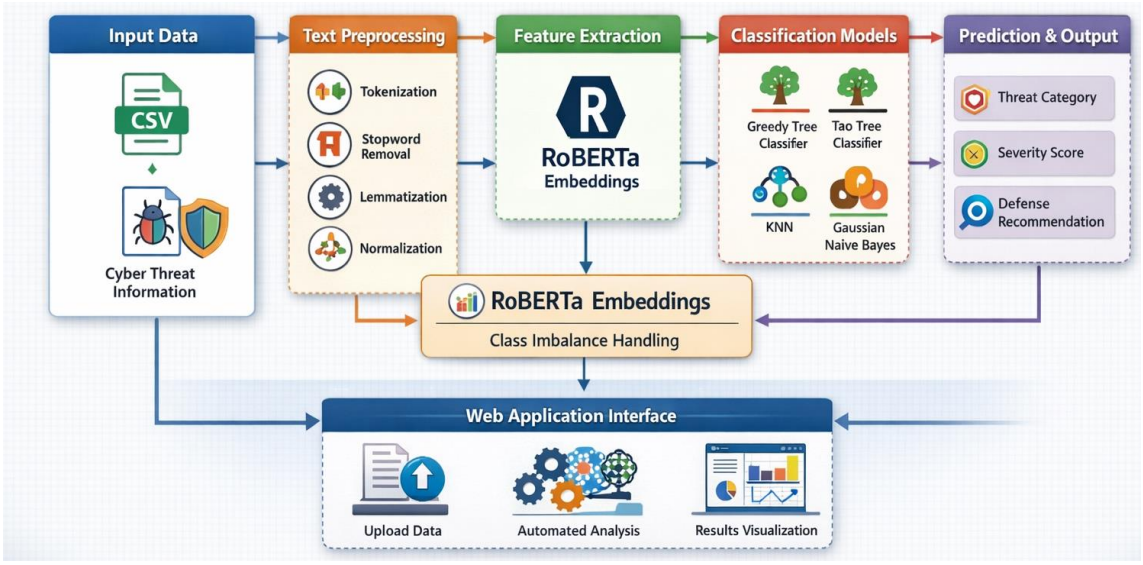

Fig 2. Proposed system architecture of medical cyber security

The system begins with the data acquisition phase, where the dataset (Medical_Cybersecurity_Dataset.csv) is loaded into a panda DataFrame. This dataset contains unstructured textual descriptions of cyber threats along with labeled attributes such as Threat Category, Severity Score, and Suggested Defense Mechanism. Initial validation checks are performed to ensure dataset integrity, establishing it as the foundation for subsequent processing. The textual data undergoes preprocessing to standardize and clean the input. This includes converting text to lowercase, tokenization, removal of stop words, and lemmatization using WordNet. At the same time, the categorical target variables are encoded into numerical form using LabelEncoder. The processed text is stored as input features, while the encoded labels are maintained along with their corresponding encoders for future inverse transformation. In this stage, the preprocessed text is transformed into dense numerical vectors using the RoBERTa Transformer model. These embeddings capture contextual meaning, semantic relationships, and linguistic nuances that traditional approaches fail to identify. The embeddings are aggregated into fixed-size feature vectors for each document and stored efficiently to support the training process. To handle class imbalance, the feature vectors are balanced using SMOTE-based resampling, generating synthetic samples for minority classes. This ensures uniform distribution across all target categories. The balanced dataset is then divided into training and testing subsets using stratified sampling, preserving class proportions in both sets.

Multiple classification algorithms, including KNN, GNB, TTC, and GTC, are trained independently. This comparative evaluation enables selection of the most effective model. The GTC model is chosen as the final model due to its superior accuracy and interpretability. All trained models are saved using joblib for future use. The trained models are evaluated on test data using standard performance metrics such as Accuracy, Precision, Recall, and F1-Score. Visualization tools are used to represent model performance. During inference, new input data follows the same preprocessing and feature extraction pipeline. The selected GTC model generates predictions for threat category, severity score, and suggested defense mechanisms. Finally, the numerical outputs are converted back into human-readable labels using inverse transformation, producing the final prediction results.

GTC model

GTC is designated as the Proposed System for analyzing healthcare cyber threats. While the TAOTree uses global optimization, the GTC is utilized for its exceptional speed and "Human-in-the-Loop"

transparency. It constructs a hierarchical decision model by making the most optimal local split at each step shown in Figure. 4.6. In the healthcare ecosystem—where rapid response to a potential data breach is critical this model provides an immediate, rule-based map that explains exactly why a specific threat (like a Phishing attempt or Malware) was flagged, based on the semantic features extracted by RoBERTa.

1. **Local Optimal Splitting (Recursive Partitioning):** The model begins at the root node, containing the entire set of RoBERTa feature vectors. It searches through every dimension of the embedding to find a single feature and a threshold that best separates the target classes (e.g., separating "DDoS" from "Ransomware"). It uses criteria like Gini Impurity or Information Gain to ensure that the resulting child nodes are as "pure" as possible.

2. **Phase 1: Feature Importance Evaluation:** Unlike black-box models, the GTC identifies which specific components of the RoBERTa vector are most influential. In our healthcare context, certain "dimensions" of the embedding might consistently represent keywords like "patient records" or "unauthorized access." The GTC prioritizes these high-impact features at the top of the tree for maximum efficiency.

3. **Phase 2: Depth-Limited Growth and Pruning:** To prevent the model from simply memorizing the training data (overfitting), the classifier employs a pruning strategy. It stops growing branches when the gain in accuracy becomes negligible or when a predefined tree depth is reached. This ensures the model remains "shallow" enough for a security analyst to read manually while remaining robust enough to handle new, unseen cyber threats.

4. **Phase 3: Multi-Target Leaf Assignment:** As the RoBERTa features trickle down through the nodes, they eventually reach a terminal "leaf." Each leaf in the proposed system represents a specific classification for Threat Category, Severity, and Defense Mechanism. The model calculates the majority class within that leaf to provide the final prediction rendered in the research's Django dashboard.

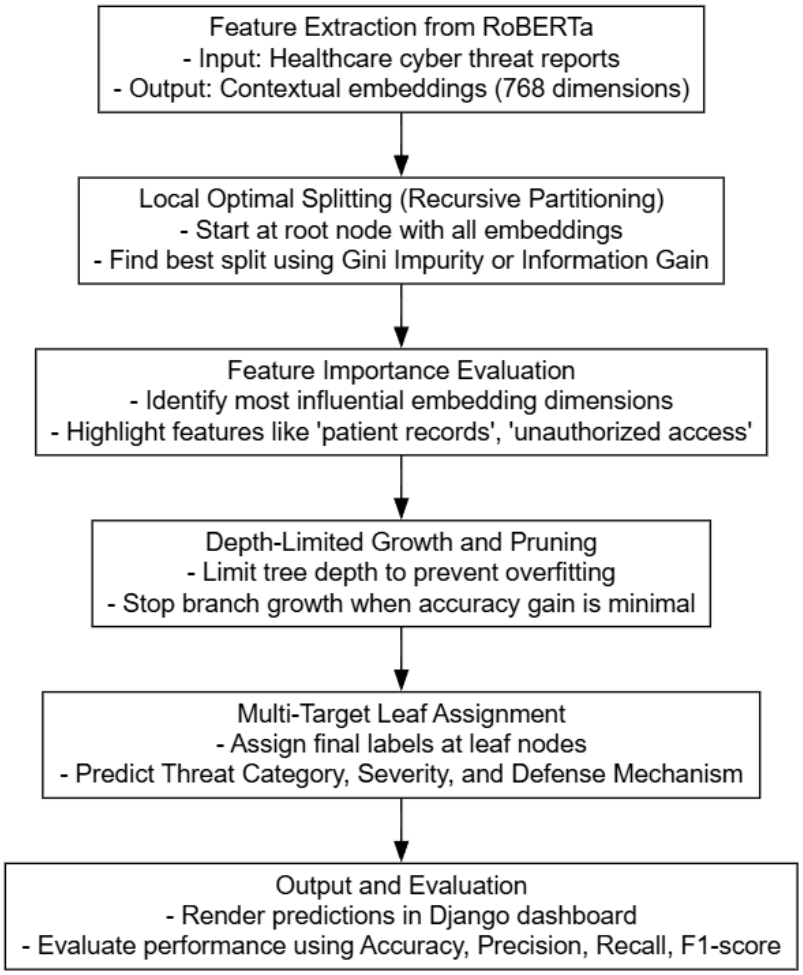

Fig 3. Internal workflow of GTC.

- **No Feature Scaling Required:** The model is invariant to the scale of the RoBERTa vectors, meaning it can process the raw output of the transformer model without needing additional normalization or standardization steps.
- **Actionable Forensics:** In the event of a cyber-attack, the tree structure acts as a forensic tool, helping security teams understand the "semantic triggers" that defined the attack's severity and category.

4. RESULT ANALYSIS

Fig. 4 presents the count distribution of samples across the target variables, providing insight into class balance within the dataset. The distribution of the Threat Category variable shows that data points are fairly evenly allocated among the four classes, although classes 1 and 2 exhibit slightly higher frequencies compared to classes 0 and 3, indicating a minor skew but overall balanced representation. Similarly, the Suggested Defense Mechanism variable demonstrates a uniform distribution across its four classes, with only negligible differences in sample counts, which is favorable for unbiased model learning. In the case of the Severity Score, the dataset is divided into five classes with moderate variation, where class 1 has the highest number of instances and class 2 has comparatively fewer samples; however, the variation is not significant enough to introduce critical imbalance issues. Overall, the distributions across all target variables reflect a well-maintained class balance, which supports stable training, minimizes bias, and enhances the reliability of predictive modeling.

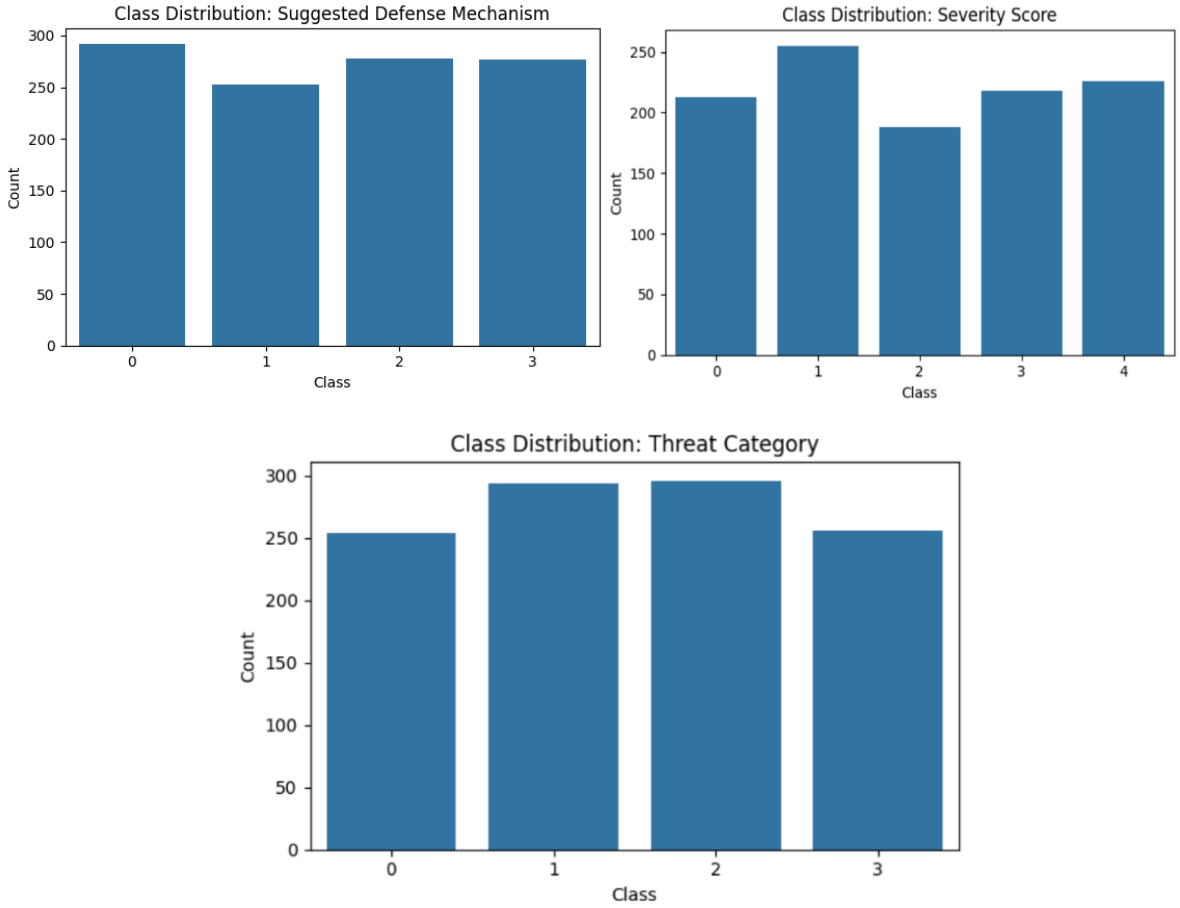

Fig 4: Count plot for all 3 target classes

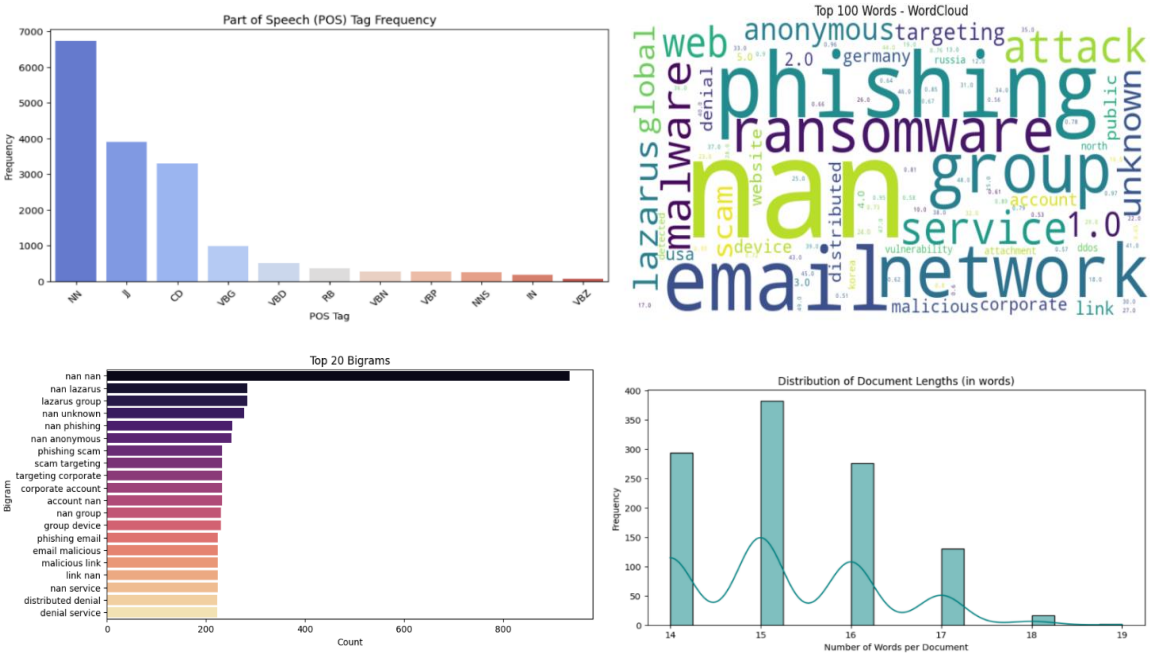

Fig 5. NLP EDA plots for different types of features in the dataset

Fig 5. shows the visualization panel provides an overview of the linguistic characteristics of the dataset, combining bigram frequency, keyword distribution, text-length variation, and POS-tag patterns to better understand the nature of the threat-intelligence text. The Top 20 Bigrams plot

highlights frequently co-occurring word pairs such as “phishing email,” “ransomware attack,” and “corporate network,” indicating recurring cyber-attack patterns and contextual themes in threat descriptions. The word cloud reinforces this by showing dominant terms like phishing, ransomware, malware, network, Lazarus group, and email, suggesting that email-based threats, targeted attacks, and advanced threat groups commonly appear in the dataset. The document length distribution plot shows that most threat descriptions fall within a narrow word-range, reflecting concise but information-rich reporting formats. Finally, the Part-of-Speech (POS) frequency chart reveals high occurrences of nouns and verbs, consistent with technical incident narratives that describe entities, actions, and attack behaviors. These plots collectively demonstrate that the dataset is semantically focused, security-contextual, and well-structured for NLP-based threat analysis and modeling.

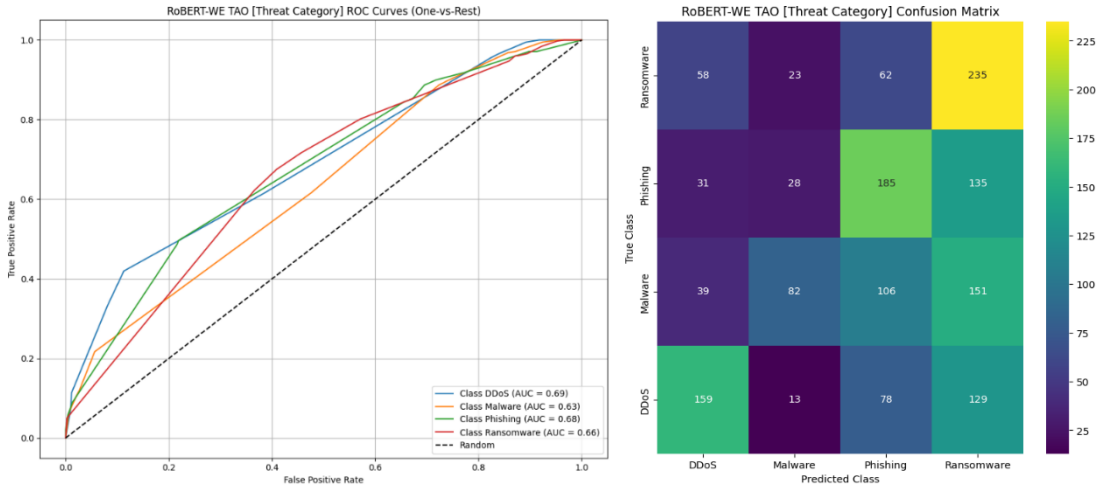

Fig 6. Confusion matrix and ROC Curve obtained for Threat Category class using TTC Model

Fig 6 shows the confusion matrix and ROC curve illustrate the classification performance of the TTC model for the Threat Category prediction task. The confusion matrix shows that while the model correctly identifies a considerable number of samples within each class, noticeable misclassifications occur between categories such as Ransomware and Phishing, and between Malware and DDoS, indicating overlapping textual patterns across attack types. Despite these confusions, the ROC curves reveal reasonably good discriminative capability across all four classes, with AUC values ranging approximately between 0.65 and 0.69, demonstrating moderate yet consistent predictive strength in distinguishing threats under a one-vs-rest evaluation. The results indicate that the TTC model performs reliably, though further improvement could be achieved through deeper representation learning or enhanced feature modeling to reduce inter-class ambiguity.

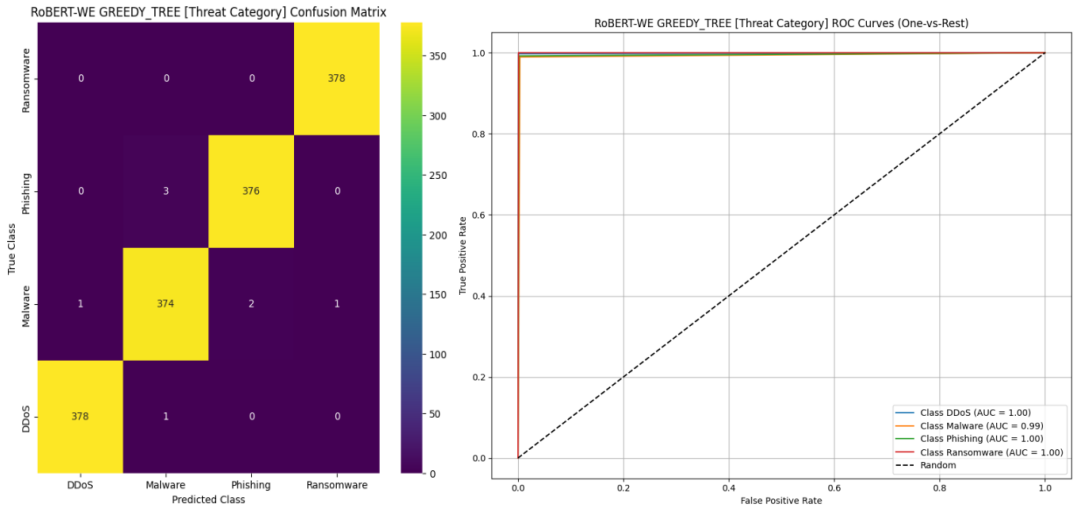

Fig 7. Confusion matrix and ROC Curve obtained for Threat Category class using GTC Model.

Fig 7. details the performance evaluation of the RoBERT-WE GTC model for classifying Threat Categories, utilizing a confusion matrix and One-vs-Rest Receiver Operating Characteristic (ROC) curves to assess predictive accuracy. The ROC analysis indicates near-perfect discriminative power, with the DDoS, Phishing, and Ransomware classes achieving a maximal Area Under the Curve (AUC) of 1.00, while the Malware class attained an exceptional AUC of 0.99. This high-performance metric is substantiated by the confusion matrix, which reveals a dense diagonal of true positive predictions; specifically, the model correctly identified 378 instances for both DDoS and Ransomware, 376 for Phishing, and 374 for Malware. The extremely sparse off-diagonal entries—such as merely 3 Phishing instances misclassified as Malware—demonstrate the model's superior precision and robustness in effectively isolating distinct threat signatures with negligible error rates.

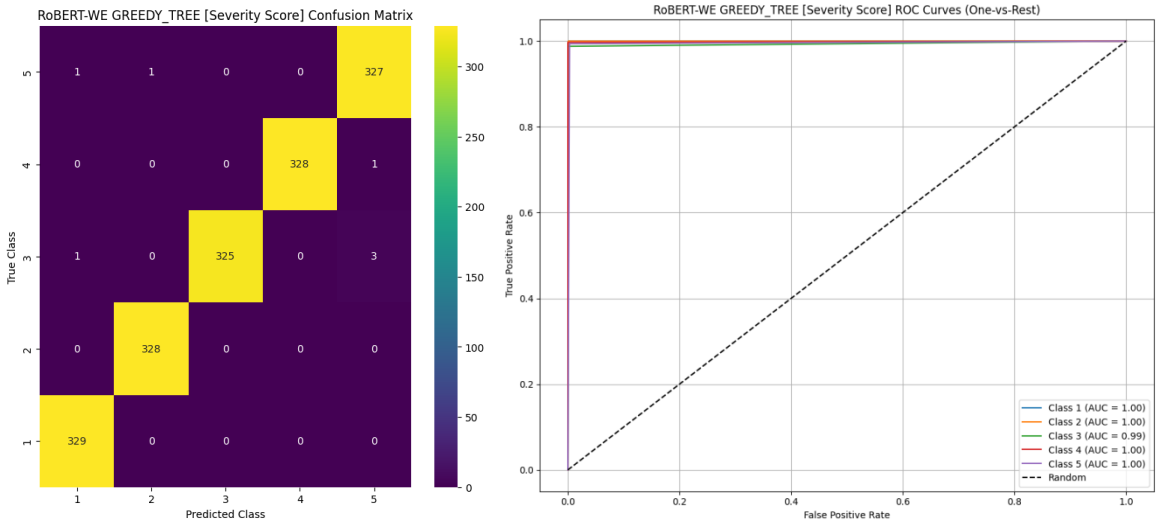

Fig 8. Confusion matrix and ROC Curve obtained for Severity Score class using GTC Model.

Fig 8 presents a comprehensive performance evaluation of the RoBERT-WE GTC model in classifying severity scores, utilizing a confusion matrix and One-vs-Rest Receiver Operating Characteristic (ROC) curves to quantify predictive accuracy. The ROC analysis demonstrates exceptional discriminative capability, with Classes 1, 2, 4, and 5 achieving a perfect Area Under the Curve (AUC) of 1.00, while Class 3 attained a near-perfect AUC of 0.99. This superior performance

is mirrored in the confusion matrix, which exhibits a highly concentrated diagonal of correct predictions; specifically, the model successfully identified 329 instances for Class 1, 328 for Class 2, 325 for Class 3, 329 for Class 4, and 327 for Class 5. The scarcity of off-diagonal entries, with only isolated misclassifications such as a single instance of Class 4 being mislabeled as Class 5, confirms the model's precision and robustness in effectively distinguishing between all severity levels with negligible error.

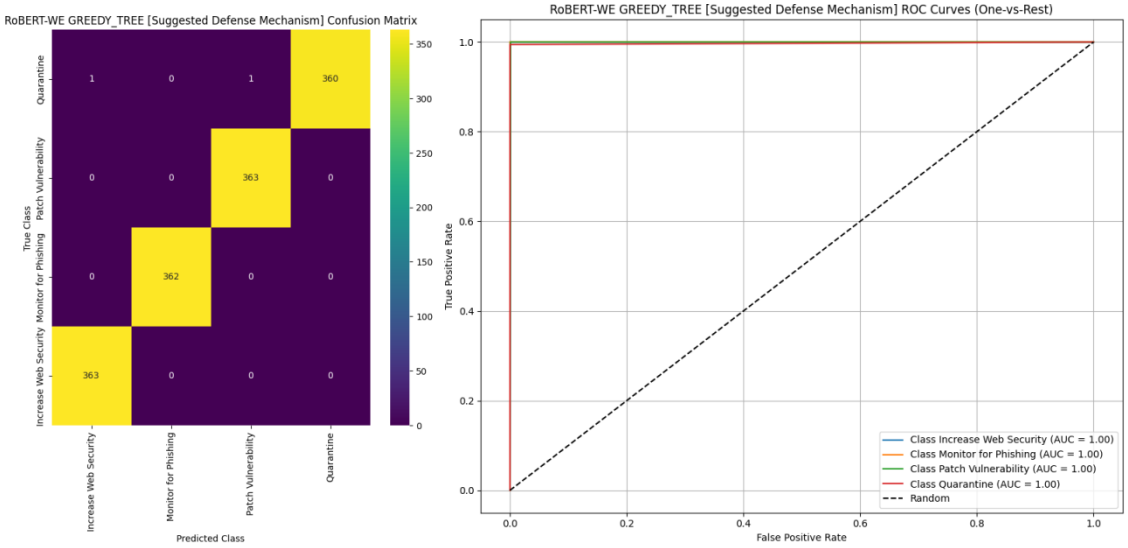

Fig 9. Illustration of confusion matrix obtained using proposed GTC model.

Fig 9. provides a technical performance assessment of the RoBERT-WE GTC model in predicting Suggested Defense Mechanisms, utilizing a confusion matrix and One-vs-Rest Receiver Operating Characteristic (ROC) curves to evaluate classification accuracy. The ROC analysis reveals perfect discriminative capability, with every defense category—Increase Web Security, Monitor for Phishing, Patch Vulnerability, and Quarantine—achieving an ideal Area Under the Curve (AUC) of 1.00. This exemplary performance is confirmed by the confusion matrix, which displays a distinct diagonal concentration of correct predictions; specifically, the model accurately identified 363 instances for Increase Web Security, 362 for Monitor for Phishing, 363 for Patch Vulnerability, and 360 for Quarantine. The model exhibits negligible error, with only two isolated misclassifications observed in the Quarantine category, demonstrating the GTC classifier's exceptional precision and reliability in automating defense mechanism recommendations.

5. CONCLUSION

An intelligent framework is introduced to enhance cybersecurity analysis in healthcare environments by leveraging advanced natural language processing techniques integrated with explainable machine learning models. The approach utilizes transformer-based embedding, specifically through the RoBERTa architecture, to convert unstructured textual data related to cyber threats into rich semantic representations suitable for downstream classification. To ensure both high performance and interpretability, the system employs tree-based models such as GTC and TTC, while GNB is incorporated as a benchmark for performance comparison. A well-defined preprocessing pipeline including text normalization, token segmentation, lemmatization, and noise filtering enhances the reliability and consistency of the input data. The entire framework is implemented within a Django-based web platform, enabling streamlined data handling, automated threat analysis, and clear visualization of outcomes for end users. Beyond classification, the system provides severity

assessment and actionable mitigation recommendations, supporting proactive defense strategies. By effectively handling challenges such as heterogeneous data and class imbalance through optimized feature engineering and scalable model design, the proposed solution demonstrates strong potential in improving cyber threat detection, strengthening decision-making processes, and enhancing the overall security posture of healthcare infrastructures.

REFERENCES

[1] Ghaffarian, S.M.; Shahriari, H.R. Software Vulnerability Analysis and Discovery Using Machine-Learning and Data-Mining Techniques: A Survey. ACM Comput. Surv. 2017, 50, 56.

[2] Satyapanich, T.; Ferraro, F.; Finin, T. CASIE: Extracting Cybersecurity Event Information from Text. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, the Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, New York, NY, USA, 7–12 February 2020; AAAI Press: Palo Alto, CA, USA, 2020; pp. 8749–8757.

[3] Gao, C.; Zhang, X.; Liu, H. Data and knowledge-driven named entity recognition for cyber security. Cybersecurity 2021, 4, 9.

[4] Nikoloudakis, Y.; Kefaloukos, I.; Klados, S.; Panagiotakis, S.; Pallis, E.; Skianis, C.; Markakis, E.K. Towards a Machine Learning Based Situational Awareness Framework for Cybersecurity: An SDN Implementation. Sensors 2021, 21, 4939.

[5] Singh, K.; Grover, S.S.; Kumar, R.K. Cyber Security Vulnerability Detection Using Natural Language Processing. In Proceedings of the 2022 IEEE World AI IoT Congress (AIIoT), Seattle, WA, USA, 6–9 June 2022; pp. 174–178.

[6] Ma, P.; Jiang, B.; Lu, Z.; Li, N.; Jiang, Z. Cybersecurity named entity recognition using bidirectional long short-term memory with conditional random fields. Tsinghua Sci. Technol. 2021, 26, 259–265.

[7] Zong, S.; Ritter, A.; Mueller, G.; Wright, E. Analyzing the Perceived Severity of Cybersecurity Threats Reported on Social Media. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; Volume 1, pp. 1380–1390.

[8] Boyd, D.; Crawford, K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Inf. Commun. Soc. 2021, 15, 662–679.

[9] Zhou, S.; Liu, J.; Zhong, X.; Zhao, W. Named Entity Recognition Using BERT with Whole World Masking in Cybersecurity Domain. In Proceedings of the 2021 IEEE 6th International Conference on Big Data Analytics (ICBDA), Xiamen, China, 5–8 March 2021; pp. 316–320.

[10] Yeboah-Ofori, A.; Mouratidis, H.; Ismai, U.; Islam, S.; Papastergiou, S. Cyber Supply Chain Threat Analysis and Prediction Using Machine Learning and Ontology. In Proceedings of the Artificial Intelligence Applications and Innovations—17th IFIP WG 12.5 International Conference, AIAI 2021, Crete, Greece, 25–27 June 2021; Springer: Cham, Switzerland, 2021; Volume 627, pp. 518–530.

[11] S. Memon, S. Memon, L. Das, B. R. Memon, Cyber security risk assessment methods for smart healthcare, in: 2024 IEEE 1st Karachi Section Humanitarian Technology Conference (KHI-HTC), 2024, pp. 1– 6. doi:10.1109/KHI-HTC60760.2024.10481961.

[12] E. Aghaei, X. Niu, W. Shadid, E. Al-Shaer, Secure BERT: A domain-specific language model for cybersecurity, in: Security and Privacy in Communication Networks, Springer, Cham, 2023, pp. 39–56

[13] K. Ameri, M. Hempel, H. Sharif, J. Lopez Jr., K. Perumalla, Cybert: Cybersecurity claim classification by fine-tuning the bert language model, Journal of Cybersecurity and Privacy 1 (2021) 615– 637.