Smart Healthcare Security: An NLP-Based Framework for Cyber Threat Prediction and Defense

S. Krishna Reddy[1], Kachu Akshaya[1], Baki Shivarama Raju[1], Polampelli Prathyusha[1], Abdullah Ameen Mohammed Hamid[1]

[1]Department of Computer Science and Engineering, [1]Sree Dattha Institute of Engineering and Science, Nagarjuna Sagar Road, Sheriguda, Ibrahimpatnam, Rangareddy Dist, 501510, Telangana, India.

To Cite this Article

S. Krishna Reddy, Kachu Akshaya, Baki Shivarama Raju, Polampelli Prathyusha, Abdullah Ameen Mohammed Hamid, "Smart Healthcare Security: An NLP-Based Framework for Cyber Threat Prediction and Defense", Journal of Science Engineering Technology and Management Science, Vol. 03, Issue 06, June 2026, pp: 650-659, DOI: http://doi.org/10.64771/jsetms.2026.v03.i06.pp650-659

Submitted: 08-05-2026 *Accepted: 15-06-2026* *Published: 22-06-2026*

ABSTRACT

The rapid digitalization of healthcare, driven by the adoption of Electronic Health Records (EHRs), connected medical devices, and advanced healthcare software systems, has significantly improved service delivery and operational efficiency. However, this increased reliance on digital infrastructure has also expanded the scope of cyber threats in the healthcare sector. Vulnerabilities present in both modern and legacy systems remain a major source of these threats, making it essential to identify and mitigate risks associated with interconnected medical devices and ICT-based healthcare environments. Threat and vulnerability analysis plays a vital role in reducing potential risks, yet it is challenging due to the large volume of unstructured data, which complicates the identification of meaningful security patterns. To address this challenge, this study presents a Natural Language Processing (NLP)-based framework for automated cyber-threat analysis using machine learning and transformer-based feature extraction. Textual threat reports are preprocessed through tokenization, lemmatization, stop-word elimination, and normalization. Contextual semantic features are then extracted using Decoding-enhanced BERT with Disentangled Attention (DeBERTa) embeddings. The generated feature set is evaluated using multiple classifiers, including Greedy Tree Classifier (GTC), Tree-based Algorithm Optimization (TAO) Tree Classifier, K-Nearest Neighbors (KNN), and Gaussian Naive Bayes (GNB), enabling comparative analysis of different learning approaches. The GTC model is selected as the optimal predictor to determine threat category, severity level, and recommended defense mechanisms. The framework is deployed as a web-based system supporting secure data upload, automated prediction, and result visualization, enhancing decision-making and enabling proactive cyber-risk management in healthcare systems.

Keywords: Cybersecurity in Healthcare, Natural Language Processing (NLP), Machine Learning, Transformer Models, DeBERTa, Electronic Health Records (EHR), Greedy Tree Classifier (GTC).

1. INTRODUCTION

Digitization in the healthcare system provides many benefits including efficiency of the healthcare service delivery, cost-savings, patient safety and care quality. There is no doubt about the positive impact of digital transformation in the healthcare sector. However, despite these benefits, the adoption of digital technology provides many Cyber Security (CS) challenges that can pose any potential risks within the healthcare system. This massive technological transformation increases the attack surface where threat actors can exploit possible threats for any potential risk within the Health Care Information Infrastructure (HCII). In recent years, several successful CS attacks were reported in the healthcare

sector: nearly 90% of healthcare organizations experienced a data breach in 2018. There are significant numbers of connected devices within the healthcare system and vulnerabilities within these connected devices can propagate to other parts of the network. An example are flaws found in Braun's infusion pump or Medtronic insulin pump, that could pose potential threat to the patient health, or simulated attacks realized to pacemakers and implantable cardiac defibrillators. Medical Internet of Things (IoT) devices are currently considered critical vulnerabilities and sources of threats and risks in the healthcare domain. Furthermore, human factors have a crucial impact on the CS within Healthcare Organizations. For these reasons, there is a need to understand the threats and vulnerabilities within the healthcare system so that control actions can be identified to ensure security of the system. However, analyzing threats and vulnerabilities in the healthcare sectors is a challenging task, due to the large number of published vulnerabilities and the difficulty in identifying the text that relates with potential threats within a healthcare system context. The large amount of unstructured Natural Language (NL) Cyber Security (CS) data related to the healthcare domain is often freely available on the Internet. More in detail, this textual data contains crucial and updated information related to the assets of the Healthcare Information Infrastructure (HCII) including threats, vulnerabilities, attacks, and other important CS information, which could be very useful to improve the protection of the HCIIs. It is often difficult to identify and extract the relevant information from such kinds of texts, which are usually available on blog posts, CS news websites, social media and other similar sources. In particular, the complexity of the NL can present polysemy, irony, long sentences and other issues, in addition to the peculiarities the technical language used CS domain, which uses many non-standard abbreviations or acronyms.

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 is an advanced Machine Learning pipeline designed for the NLP-Based Analysis of Cyber Threats and Vulnerabilities in the Healthcare Ecosystem. It starts by preprocessing raw threat text using NLTK for cleaning and lemmatization, while simultaneously encoding the three target variables (Threat Category, Severity Score, Defense Mechanism). The core innovation lies in the use of the DeBERTa Transformer model to generate dense, context-aware embeddings from the text, effectively capturing semantic meaning that traditional systems missed. After balancing the imbalanced dataset using SMOTE, the rich DeBERTa features are fed into multiple supervised classifiers, primarily the GTC (Proposed Model), which is chosen for its balance of high predictive accuracy and crucial model interpretability, allowing the system to provide fast, accurate, and explainable threat classifications for actionable defense strategies as demonstrate in figure. 1.

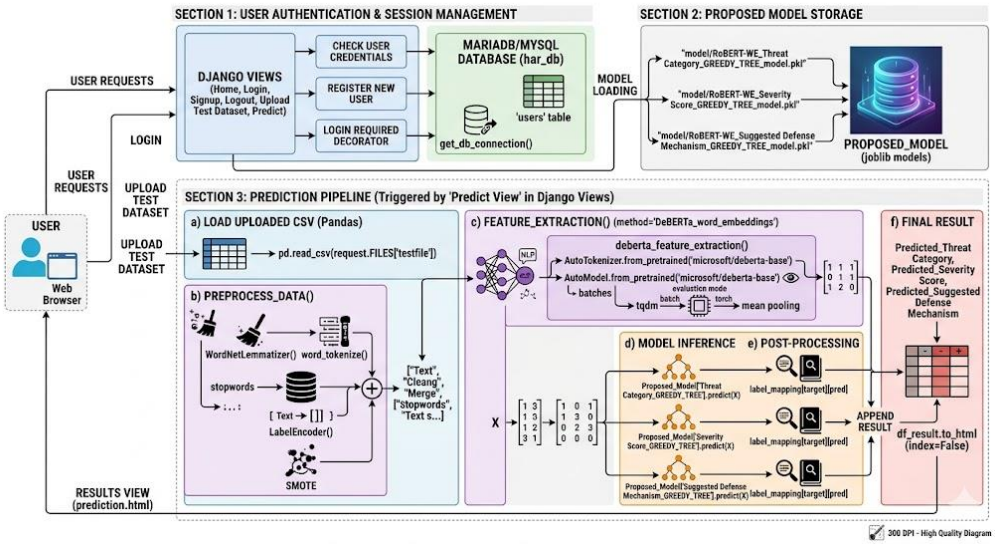


Figure 1: Proposed system architecture

1. Data Acquisition and Dataset Loading

The system initiates with the Data Acquisition step, where the core dataset (Medical_Cybersecurity_Dataset.csv) is loaded into a pandas DataFrame using the upload_dataset function. This dataset contains the raw, unstructured text descriptions of cyber threats and the corresponding human-labeled categorical columns: Threat Category, Severity Score, and Suggested Defense Mechanism. The integrity of the dataset is checked immediately, and it is prepared as the primary source of truth for all subsequent NLP and ML tasks.

2. Text Preprocessing and Label Encoding

Once loaded, the text data undergoes extensive Text Preprocessing to standardize the linguistic input. This critical clean_text pipeline involves lower-casing the text, performing tokenization, removing common English stop words (nltk.corpus.stopwords), and applying WordNet Lemmatization to reduce words to their dictionary roots (e.g., 'hacking' to 'hack'). Simultaneously, the three categorical target columns are separated and converted to numerical integers via LabelEncoder. The processed text is then consolidated into a single input column (X), while the encoded labels (Y_dict) are stored alongside the original encoders for later inverse transformation.

3. Deep Contextual Feature Extraction (DeBERTa Embeddings)

This step leverages the state-of-the-art DeBERTa Transformer model for feature engineering. The preprocessed text is input into DeBERTa, which converts the documents into high-dimensional, dense numerical vectors known as embeddings. This process captures the deep semantic meaning, context, and relationships between words, allowing the models to understand nuances that traditional keyword-based systems miss. The embeddings are aggregated, typically using mean pooling, to create a single fixed-size feature vector (X) for each text document, which is then cached to ensure training efficiency.

4. Data Balancing and Stratified Train/Test Split

To counteract the inherent class imbalance in cyber security data, a critical balancing technique is applied. The DeBERTa feature vectors are simultaneously oversampled and then subjected to SMOTE (Synthetic Minority Over-sampling Technique), which generates synthetic data points for the minority classes, ensuring the training data is balanced for each of the three classification tasks. Following balancing, the data is split into 80% training and 20% testing sets using train_test_split with stratification, which ensures that the class proportions are maintained in both the training and testing datasets.

5. Multi-Algorithm Model Training and Selection

Multiple machine learning classifiers, including KNN, GNB, and two Interpretable Decision Trees (TAO Tree and GTC), are trained independently for each of the three targets. This comparative approach allows for the selection of the best model. Since GTC (Proposed Model) offers both high performance and the crucial benefit of interpretability (essential for security analysts), it is highlighted for deployment. Trained models and their configurations are saved using joblib.

6. Performance Evaluation and Real-World Inference

The trained models are rigorously evaluated on the unseen test data. A custom Metrics Calculator class computes standard performance metrics (Accuracy, Precision, Recall, F1-Score) and a Graph Plotter visualizes the results. Finally, the selected GTC is used to perform inference on new test data, which must pass through the exact same feature extraction pipeline. The numerical predictions are then immediately converted back into human-readable labels using the stored inverse transform function of the original Label Encoder objects, producing an actionable final report.

GTC model

The GTC (implemented via the imodels library) is designated as the Proposed System for analyzing healthcare cyber threats. While the TAO Tree 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. 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 DeBERTa as shown in figure. 2.

1. **Local Optimal Splitting (Recursive Partitioning):** The model begins at the root node, containing the entire set of DeBERTa 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 GTCidentifies which specific components of the DeBERTa vector are most influential. In our healthcare context, certain "dimensions" of the embedding might consistently represent keywords like "patient records" or "unauthorized access." The GTCprioritizes 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 DeBERTa 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 project's Django dashboard.

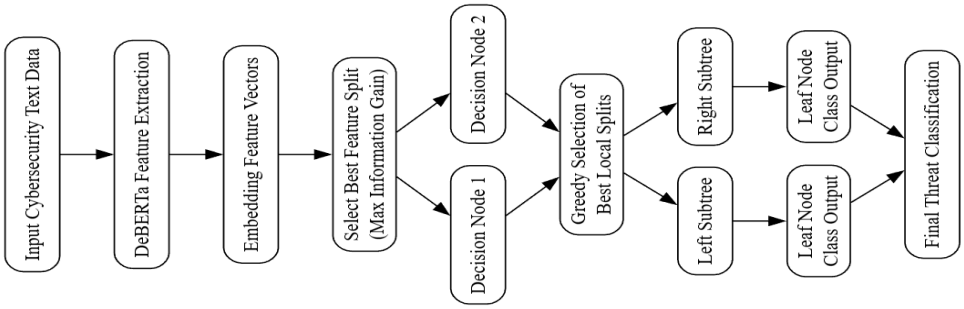


Figure 2: Internal workflow of GTC

4. Result Analysis

The figure 3 shows count-distribution plots illustrate the balance of samples across the target labels in the dataset, showing how instances are distributed for Threat Category, Suggested Defense Mechanism, and Severity Score. The Threat Category plot indicates a relatively even spread of records across four classes, with classes 1 and 2 having slightly higher representation compared to classes 0 and 3, suggesting a near-balanced but mildly skewed dataset. The Suggested Defense Mechanism distribution also appears balanced across four classes, with only small variations in class frequency, which supports fair model learning without strong class dominance. Meanwhile, the Severity Score plot shows five classes with moderately varied counts, where class 1 has the highest frequency and class 2 has the lowest, but the difference is not extreme enough to cause severe imbalance. Overall, the distributions across all three target variables suggest that the dataset is generally well-balanced, enabling reliable model training and reducing the likelihood of bias toward any particular class during prediction

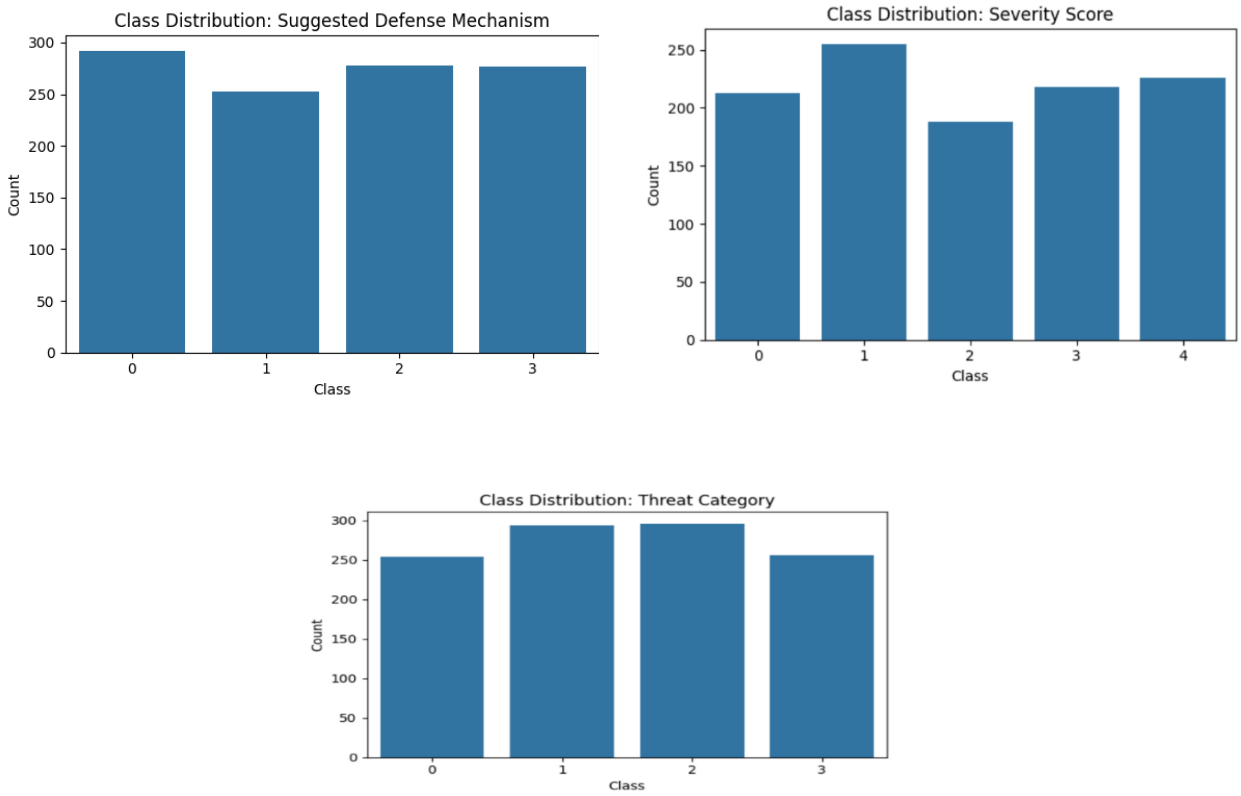


Figure 3: Count plots for all the three categories

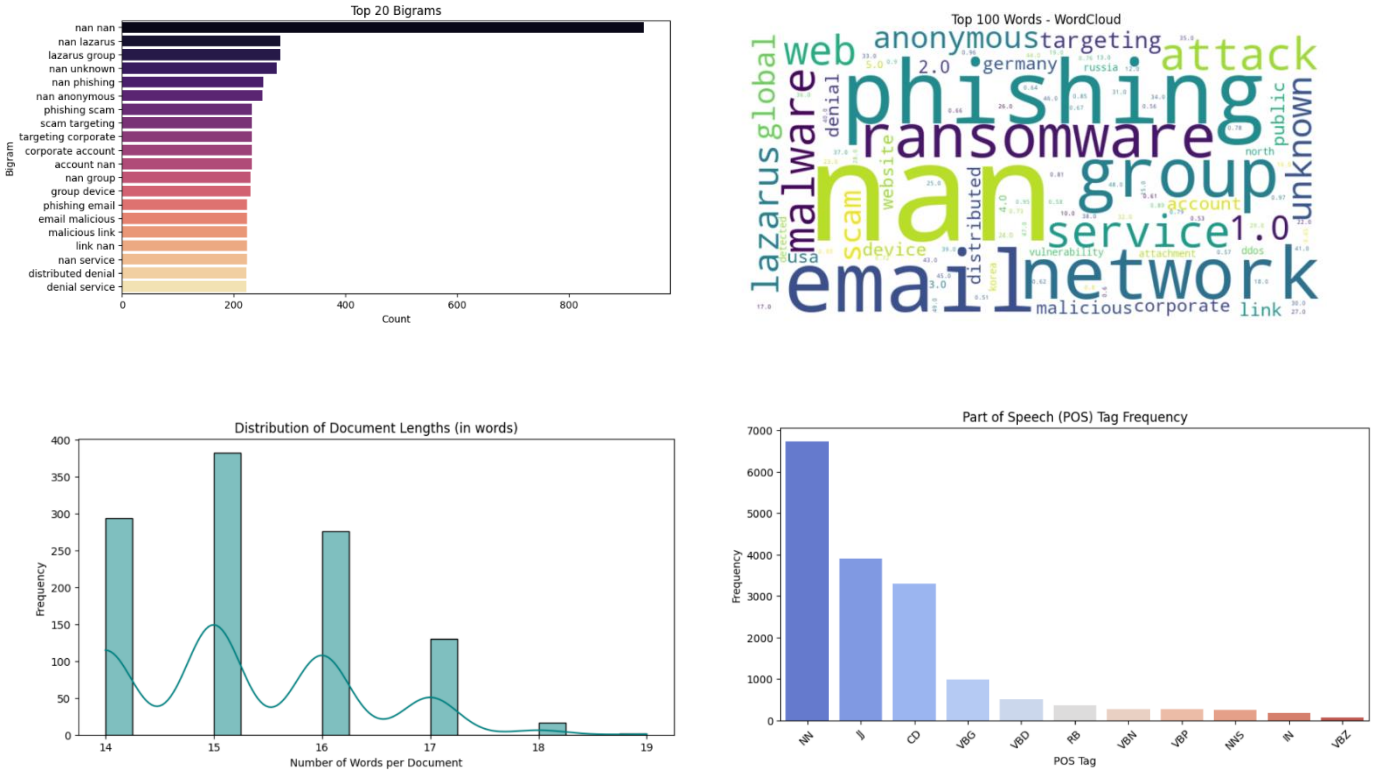


Figure 4: EDA plots

The figure 4 shows 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.

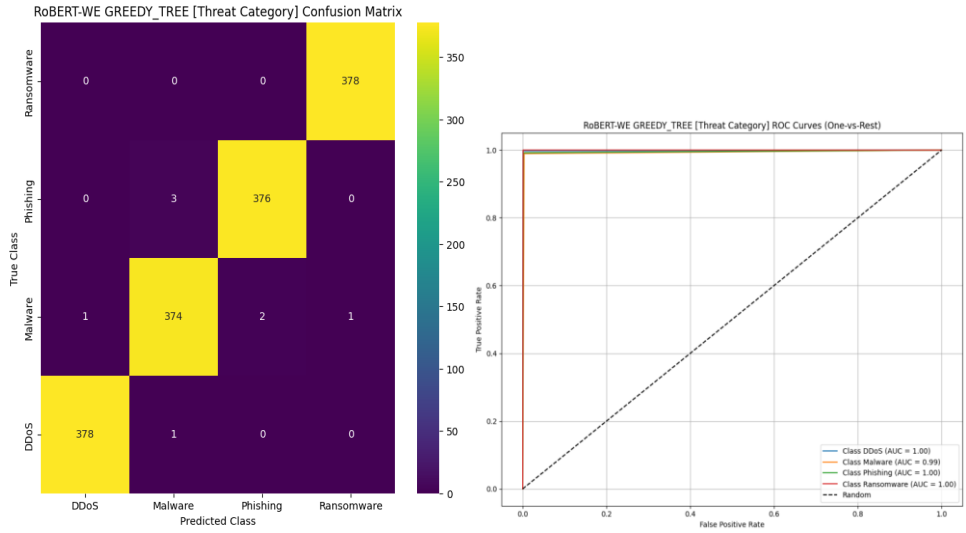


Figure 5: Confusion matrix and ROC Curve obtained for Threat Category class using GTC Model.

The figure 5 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.

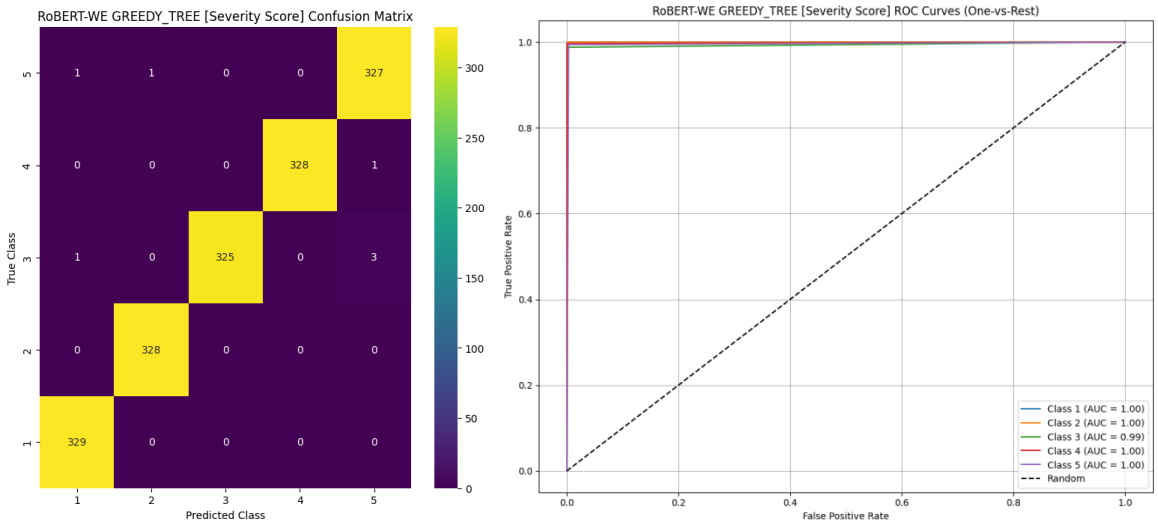


Figure 6: Confusion matrix and ROC Curve obtained for Severity Score class using GTC Model.

The figure 6 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.

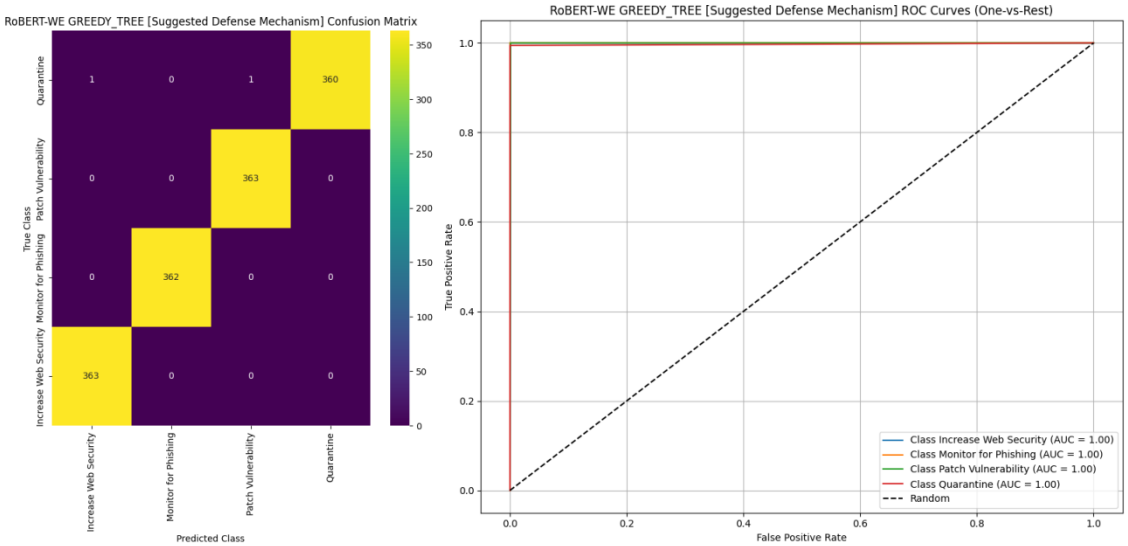
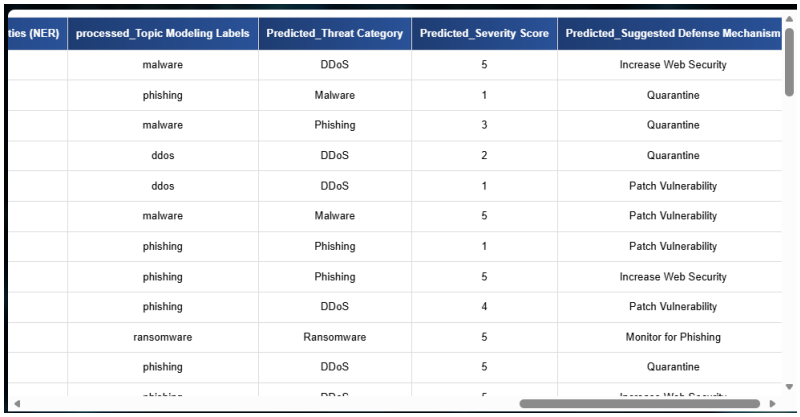


Figure 7: Confusion matrix and ROC Curve obtained for Suggested Defense Mechanism class using GTC Model.

The figure 7 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.



ties (NER)	processed_Topic Modeling Labels	Predicted_Threat Category	Predicted_Severity Score	Predicted_Suggested Defense Mechanism
	malware	DDoS	5	Increase Web Security
	phishing	Malware	1	Quarantine
	malware	Phishing	3	Quarantine
	ddos	DDoS	2	Quarantine
	ddos	DDoS	1	Patch Vulnerability
	malware	Malware	5	Patch Vulnerability
	phishing	Phishing	1	Patch Vulnerability
	phishing	Phishing	5	Increase Web Security
	phishing	DDoS	4	Patch Vulnerability
	ransomware	Ransomware	5	Monitor for Phishing
	phishing	DDoS	5	Quarantine

Figure. 8: Predictions on sample test data.

figure 8 presents the prediction results generated by the proposed NLP-driven cyber-threat analysis framework on sample healthcare-related test data, illustrating the complete workflow from input to output. The figure shows raw textual threat reports being processed through preprocessing steps such as tokenization, lemmatization, stop-word removal, and normalization, followed by contextual feature extraction using DeBERTa embeddings. These features are fed into the trained prediction engine, where the GTC produces three key outputs for each input instance: the threat category, identifying the type of cyber attack the severity score, indicating the level of risk associated with the threat; and the suggested defense mechanism, providing appropriate mitigation strategies. The results are displayed through a web-based interface in a structured and interpretable format, enabling efficient analysis and supporting cybersecurity decision-making within the healthcare environment.

5. CONCLUSION

This study presents a robust NLP-driven framework for the detection and evaluation of cyber threats and vulnerabilities within healthcare systems, combining advanced transformer-based feature extraction with interpretable machine learning techniques. The use of the DeBERTa model enabled effective generation of contextual embeddings, capturing complex semantic relationships in cybersecurity text data and transforming them into meaningful numerical representations for classification tasks. The implementation of GTC and TAO classifiers ensured strong predictive accuracy while maintaining model transparency and interpretability. In addition, the inclusion of GNB provided a probabilistic reference model for performance comparison. The preprocessing pipeline, consisting of tokenization, lemmatization, and noise reduction, played a crucial role in enhancing data quality and improving overall model performance. Furthermore, the deployment of the framework through a Django-based web application enabled efficient user interaction, secure data uploads, and real-time visualization of predictions, making the system suitable for practical cybersecurity applications. The model demonstrated the ability to accurately classify threat types, assess severity levels, and recommend appropriate defense strategies, thereby supporting informed and proactive decision-making in healthcare environments.

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.