Milestone Transactions on
Artificial Intelligence
Vol. 01 | Issue. 01 | 2026

RESEARCH ARTICLE OPEN ACCESS

Enhancing TARA Through a Continuous Improvement Method Based on Cybersecurity Threat Intelligence

M Munibabu* **. P Naga Sashank . C Bhanu . B Praveen Kumar . J Duraj Reddy**

Department of CSE (IoT, Cyber Security including Block Chain Technology),
Annamacharya Institute of Technology & Sciences (Autonomous),
Tirupati, A.P, India.

DOI: **10.5281/zenodo.18508416**
Received: 11 January 2026 / Revised: 19 January 2026 / Accepted: 6 February 2026
**Corresponding Author: marathi.muni@gmail.com*

Abstract – The ever-increasing complexity and dynamics of cyber threats demand risk assessment tools that extend beyond the realm of static and periodic risk assessments. Although Threat Analysis and Risk Assessment (TARA) techniques are systematic, they are not flexible enough to handle dynamic cyber threats. To overcome the limitations of the conventional Threat Analysis and Risk Assessment techniques, this paper proposes a continuous improvement approach that incorporates Cybersecurity Threat Intelligence (CTI) and an intelligent risk prediction model. In this paper, a hybrid multimodal ensemble approach has been proposed to predict the CVSS score based on the structured vulnerability attributes and unstructured textual threat intelligence obtained from the CVE data. Comprehensive experiments have been performed using the large-scale CVE dataset to prove the efficiency of the proposed approach in efficiently predicting the risk with an R^2 value of 0.9947 and a Mean Absolute Error of 0.0133. Explainability analysis is performed to ensure that the proposed approach meets the predefined cybersecurity risk principles. The experimental outcome clearly demonstrates that the proposed CTI-based continuous TARA approach can enhance the accuracy of risk prediction in a dynamic cybersecurity setting.

Index Terms – Cybersecurity Threat Intelligence, Threat Analysis and Risk Assessment, CVSS Prediction, Ensemble Learning, Random Forest, XGBoost, BERT, Continuous Risk Assessment

I. INTRODUCTION

Nowadays, the issue of cyber threats is no longer just about isolated incidents; rather, it is about sustained, dynamic, and intelligence-driven threats that can have serious financial, operational, and safety consequences. Therefore, cybersecurity has emerged as a strategic imperative that requires proactive, data-

driven defense strategies rather than the traditional reactive approach [1]. Cybersecurity Threat Intelligence (CTI) has emerged as an important aspect of modern cybersecurity operations, providing valuable insights into attacker behavior, enabling organizations to make informed decisions to strengthen their security posture [2]. In addition to CTI, Threat Analysis and Risk Assessment (TARA) has emerged as a basic methodology for threat identification, risk estimation, and risk mitigation planning. TARA is applied in safety-critical and cyber-physical systems and is included in security engineering standards such as ISO/SAE 21434. It offers a systematic approach to assess the likelihood and consequences of possible attacks throughout the entire system life cycle [3]. Nevertheless, traditional TARA processes are normally carried out in designated design phases or review cycles. This makes them less effective in environments where threat profiles change constantly due to newly discovered vulnerabilities, zero-day attacks, and shifting attacker tactics [4].

Integrating CTI into TARA processes could transform traditional risk assessment into a dynamic, continuously improving mechanism. By leveraging real-time threat feeds, historical attack data, and contextual intelligence, risk evaluations can shift from assumption-based estimation to evidence-driven analysis [5]. However, there are some challenges that hinder the effective integration of CTI, such as the sheer volume and diversity of CTI, noise and uncertainty in intelligence, and the absence of automated means to convert intelligence into quantitative risk updates. Manual processing of CTI is labor-intensive, error-prone, and not scalable to the speed of contemporary threat landscapes [6]. Machine learning approaches offer a potential solution to these issues, as they can be used to automatically detect patterns, make predictions, and even recalculate risk values. Among the various machine learning algorithms, the Random Forest Regressor is found to be the most suitable approach to implement the CTI-driven risk assessment, owing to its robustness, immunity to overfitting, and capacity to deal with high-dimensional and nonlinear data, along with its excellent predictive accuracy on various security-related datasets [7]. The potential of the Random Forest approach can be leveraged to effectively predict risk values and recalculate them using the TARA framework.

Accordingly, we propose the continuous improvement approach to TARA that utilizes Cybersecurity Threat Intelligence by incorporating CTI data streams into the assessment process and using a Random Forest Regressor to predict and improve risk levels. Our proposed approach will enable TARA to evolve from a periodic risk assessment approach to an adaptive and intelligence-driven approach that can react to new threats in real-time. This proposed approach seeks to improve the accuracy of risk predictions, lower the latency of risk assessments, and improve the resilience of cybersecurity risk management in complex digital environments [8].

The main contributions are as follows:

- Redefining TARA as a Living Risk Assessment Process: This research paradigm shifts TARA from a static, lifecycle-bound process to a living, intelligence-driven process that is informed by real-world Cybersecurity Threat Intelligence (CTI).
- Intelligence-to-Risk Translation Mechanism: In this research, a novel intelligence-to-risk translation mechanism is proposed that effectively translates diverse CTI into quantitative risk

updates based on the CVSS score, thus facilitating evidence-based and automated risk update processes within formal TARA.

- First Multimodal Ensemble for CTI-Integrated TARA: To the best of our knowledge, this research is the first to propose a multimodal ensemble framework (RF-XGB-BERT) that integrates structured vulnerability data and unstructured threat narratives for continuous cybersecurity risk assessment.

- Semantic Risk Awareness via Transformer-Based Threat Understanding: This research proposes a transformer-based semantic risk awareness framework that incorporates semantic understanding of vulnerability descriptions, thus enabling the framework to capture implicit risk severity information that is not captured by traditional CVSS- or rule-based risk models.

- Explainable and Analyst-Centric Risk Intelligence: The framework also supports dual-level explainability, including structural explainability based on SHAP values and token-level semantic interpretation. This guarantees that the predictions made by the model are explainable and consistent with the analyst's intuition.

- Empirical Evidence for Adaptive Cyber Risk Management: Significant evaluation on large-scale real-world CVE data sets revealed that the proposed approach not only achieves state-of-the-art accuracy but also supports low-latency, dynamically updated predictions.

II. LITERATURE SURVEY

Presently, the focus of the latest research in cybersecurity is on the need to integrate Cyber Threat Intelligence and Artificial Intelligence to create a strategy that is proactive and adaptive in its approach. The present risk assessment methodologies, such as Threat Analysis and Risk Assessment (TARA), rely on periodic assessments, which are not effective in the present scenario. This has led to the concept of automated intelligence extraction and the use of AI-based analytics. In addition, Rahmati et al. [9] presented an explainable and lightweight AI model for real-time cyber threat hunting in edge networks. The model focused on ensuring transparency and efficiency in machine learning models. This is important for risk assessment processes since model justification is critical. In addition, Sorokoletova, Antonioni, and Colò et al.[10] presented a scalable AI-driven CTI extraction model that utilizes a transformer architecture to organize unstructured intelligence feeds in a standardized manner. This is significant since it indicates how intelligence processing can be automated to improve interoperability for risk modeling purposes.

To overcome the challenges of data quality and preprocessing, Al-Yasiri et al. [11] presented a conceptual model for multilingual CTI event extraction using advanced natural language processing and sequence learning approaches. The study results confirm the significance of proper intelligence normalization prior to integration into automated cybersecurity systems. In this regard, Barakat et al. [12] presented a review of various AI-based threat intelligence mechanisms, stating that standardized intelligence sharing formats, along with machine learning approaches, are essential for developing effective cyber defense ecosystems. The development of generative and predictive types of AI has also increased the capabilities of CTI. Balasubramanian et al. [13] studied the application of generative AI in

CTI and found that the capabilities of intelligence correlation and multimodal data fusion were enhanced, but there are some issues regarding the scalability and misinformation problems. At the enterprise level, Kwentoa [14] demonstrated that the application of AI-based CTI systems can greatly improve the accuracy of detection and prioritization compared to the conventional rule-based methods, but there are some problems regarding bias and interpretability issues.

Apart from the role of intelligence extraction, predictive analytics has also emerged as a significant aspect for the transformation of CTI into actionable intelligence. Panda et al. [15] developed a predictive CTI framework that can predict emerging threats using anomaly detection and pattern recognition techniques driven by AI. Other research works on the forecasting of cyber attacks have also proved the potential of time series and deep learning-based models for the prediction of threats in the context of prioritization strategies [16]. In addition, research works on NLP-based CTI workflows have proved the potential of structured intelligence for minimizing the workload of risk assessment cycles [17]. However, despite these developments, the majority of the existing work is focused on detection, extraction, or prediction as standalone processes rather than integrating the outputs of CTI into risk assessment frameworks. There is a need for a continuous improvement approach that connects intelligence updates to the risk recalibration process in TARA. This requires predictive models that can process diverse intelligence features and update risk values dynamically.

Ensemble learning methods, especially the Random Forest Regression, are suitable for this purpose because of their robustness, capacity to handle non-linear data, and efficiency in processing high-dimensional security data. As observed in systematic reviews of CTI technology, machine learning-driven risk prediction is a major enabler of adaptive cybersecurity systems [18]. Thus, the combination of Random Forest-driven predictive modeling and CTI-integrated TARA can help fill the existing gap between intelligence gathering and continuous risk improvement.

Table 1: Overview of existing studies

Ref	Focus Area	Method/Technique Used	Key Contribution	Limitation
[9]	Real-time threat hunting	Explainable & lightweight ML models	Improves transparency and efficiency of AI-based threat detection in edge environments	Focused on detection, not risk assessment integration
[10]	CTI information extraction	Transformer-based AI framework	Automates structuring of raw CTI feeds into standardized formats	Does not connect CTI outputs with formal risk models
[11]	CTI event extraction	NLP, XLM-RoBERTa, sequence learning	Enhances preprocessing and normalization of multilingual CTI data	Limited to data preparation stage
[12]	AI-driven CTI systems	Review of AI + intelligence frameworks	Highlights importance of automation and intelligence-sharing standards	Conceptual; lacks predictive modeling
[13]	Generative AI in CTI	Generative AI & multimodal data fusion	Improves intelligence correlation and situational awareness	Scalability and misinformation risks
[14]	Enterprise CTI platforms	AI-based detection & prioritization models	Shows improved detection accuracy in enterprise security	Interpretability and bias issues remain

[15]	Predictive CTI	AI anomaly detection & pattern recognition	Enables early-warning capabilities for emerging threats	Not integrated with structured risk assessment methods
[16]	Cyber attack prediction	Time-series & deep learning models	Forecasts threat trends and supports proactive defense	Focus on forecasting, not continuous risk recalibration
[17]	CTI automation	NLP + ML pipelines	Converts unstructured reports into actionable indicators	Limited linkage to risk scoring frameworks
[18]	CTI technologies & effectiveness	Systematic review methodology	Identifies AI as key enabler of adaptive cybersecurity	Does not propose implementation model

III. METHODS & MATERIALS

A. Dataset Description

For our cybersecurity threat analysis study, we relied on the CVE (Common Vulnerabilities and Exposures) dataset from Kaggle. The CVE catalog, administered by the National Cybersecurity FFRDC with the Mitre Corporation and supported by the U.S. Department of Homeland Security, is a standardized resource for publicly available information about vulnerabilities and exposures in information security. The dataset is quite extensive, with tens of thousands of records, each associated with a distinct CVE identifier. Some of the most interesting attributes include the CVE identifier, a description of the vulnerability, the publication and last update dates, severity metrics such as CVSS scores, and the affected software or hardware products. This extensive resource makes it possible to analyze threat trends in-depth, makes it easier to find high-priority vulnerabilities, and provides the foundation for studies on automated vulnerability detection and the creation of cybersecurity policies. It has grown to be a well-liked resource for researchers and security professionals looking to comprehend, categorize, and lessen possible cyberthreats because of its well-organized structure and extensive coverage [19].

B. Data Preprocessing

We selected and preprocessed the CVE (Common Vulnerabilities and Exposures) dataset to facilitate ongoing enhancements to Threat Assessment and Risk Analysis (TARA) using cybersecurity threat information. Type safety, consistent feature representation, and preparedness for vulnerability severity score (CVSS) regression modeling were all guaranteed by the preprocessing procedure.

I. Missing value handling: The dataset includes numerical scores, structured categorical data, and unstructured text. Missing values were handled in a type-safe manner to preserve data integrity:

- Textual description x_{text}

$$x_{text} \leftarrow x_{text} \text{ if present,} \qquad \text{else empty string ("")}$$

Without adding nulls, this maintains unstructured threat intelligence.

- CVSS score (target variable) y

$$y \leftarrow \begin{cases} y, & if\ present \\ median\ (y), & if\ missing \end{cases}$$

By avoiding regression toward the mean, median imputation maintains the central tendency.

- Structured categorical columns x_{cat}

$$x_{cat} \leftarrow \begin{cases} x_{cat}, & if\ present \\ \text{UNKNOWN}, & if missing \end{cases}$$

This method keeps the category representation consistent across characteristics like impact measures, access vectors, and CWE codes.

II. Column Standardization and Feature Selection: Column names were standardized to provide semantic agreement with TARA risk modeling:

summary → description,cvss → cvss_score,pub_date → published_date

Only characteristics pertinent to CVSS regression were kept:

X = {description, cwe_code, access_authentication, …, impact_integrity}

To preserve the integrity of the dataset, entries without a target or description were eliminated.

III. Temporal Feature Engineering: A datetime format was applied to the published_date field. To record temporal trends in vulnerabilities, which are crucial for TARA's ongoing improvement and enable the system to take changing threat patterns into account, the year and month were extracted:

year = published_date.year, month = published_date.month

IV. Categorical Encoding: Label encoding was used to encode structured categorical features:

$$x_{cat} \leftarrow \text{LabelEncoder}\ (x_{cat})$$

In regression models, this numerical representation enables the incorporation of metadata alongside text features.

V. Feature Representation for Regression

a. Text Features x_{text}

TF-IDF vectorization was used to convert unstructured threat intelligence from the description field:

$$x_{text} \leftarrow \text{TF} - \text{IDF}\ (x_{text})$$

To capture significant textual patterns, we analyzed unigrams and bigrams with up to 15,000 features.

b. Structured Features x_{meta}

Text embeddings were paired with metadata elements, such as temporal features and encoded category variables:

$$\text{X} = [x_{text}, x_{meta}]$$

vi. The dataset was separated into training, validation, and test sets to assess prediction performance:

70% train,15% validation,15% test

This division validates the TARA model's capacity to generalize to unseen threats while enabling it to learn from past vulnerabilities.

C. Proposed Model Architecture

In order to accurately forecast CVSS for continuous enhancements in Threat Analysis and Risk Assessment (TARA), this work presents a Hybrid Multimodal Ensemble for CVSS Prediction (RF–XGB–

Figure 1: Graphical representation of the proposed model architecture BERT) that combines unstructured textual threat intelligence with structured vulnerability criteria. The architecture is intended to generate strong, statistically sound predictions appropriate for cybersecurity decision-making while capturing complementary risk signals from diverse data sources. Figure 1 depicts the proposed model architecture. Table 1 presents the hyperparameter configuration of the proposed model framework.

Let, a vulnerability instance be expressed as:

$$\mathcal{V}_i = \{x_i^{(s)}, x_i^{(t)}, y_i\}$$

where, $x_i^{(s)} \in \mathbb{R}^d$ define structured features, $x_i^{(t)}$ represent unstructured textual derived from threat intelligence sources, and $y_i \in \mathbb{R}$ is the ground-truth-CVSS score.
The objective is to learn a function:

$$\hat{y}_i = f(x_i^{(s)}, \quad x_i^{(t)}$$

that maintains generalization across unknown threats while reducing prediction error.

I. Structured Feature Learning: Two complementary tree-based regressors are used to model structured features, thereby capturing ensemble diversity and nonlinear interactions.

- Random Forest Regressor (RF): An ensemble of decision trees trained using bootstrapped samples is learned by Random Forest. The definition of an RF prediction is:

$$\hat{y}_i^{RF} = \frac{1}{T}\sum_{t=1}^{T} h_t(x_i^{(s)})$$

where, $h_t(x_i^{(s)}$ define the prediction of the t-th decision tree and T is the number of trees. RF is a dependable baseline for CVSS estimation because it performs well across diverse structured inputs and is robust to overfitting.

- XGBoost Regressor (XGB): Using gradient-boosted decision trees, XGBoost maximizes the following objective:

$$\mathcal{L}^{(k)} = \sum_i \ell(y_i, \hat{y}_i^{(k-1)} + f_k(x_i^{(s)} + \Omega\,(f_k)$$

where, f_k is the k-th boosting tree. The resulting prediction is:

$$\hat{y}_i^{XGB} = \sum_{k=1}^{K} f_k(x_i^{(s)})$$

By concentrating on difficult-to-predict samples and capturing fine-grained feature interactions, XGBoost enhances RF.

II. Textual Threat Intelligence Modeling: A BERT-based regression model is used to process unstructured vulnerability descriptions, allowing threat narratives to be understood semantically.

- Text Encoding: Given an input sequence that has been tokenized:

$$x_i^{(t)} = \{w_1,\ w_2, \ldots\ldots\ldots\ldots\ldots, w_n\}$$

BERT maps for every token to contextual embeddings:

$$H_i = BERT\left(x_i^{(t)}\right) \in \mathbb{R}^{n \times d_h}$$

The [CLS] token is h_{CLS} which employed for regression.

- BERT Regression Head: The CVSS score is predicted using a layer of linear regression:

$$\hat{y}_i^{BERT} = w^\top h_{CLS} + b$$

The Mean Squared Error (MSE) is used to optimize the model:

$$\mathcal{L}_{BERT} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i^{BERT})^2$$

This feature allows the model to incorporate contextual risk indicators, exploit descriptions, and implicit severity cues that are not captured in structured fields.

III. Multimodal Ensemble Fusion: We use a weighted late-fusion approach to combine predictions from both structured and unstructured modalities. The final CVSS estimate is calculated as follows:

$$\hat{y}_i^{Ens} = \alpha\hat{y}_i^{RF} + \beta\hat{y}_i^{XGB} + \gamma\hat{y}_i^{BERT}$$

subject to:

$$(\alpha, \beta, \gamma) = (0.3, 0.3, 0.4)$$

This weighting maintains structured feature reliability while reflecting the significant contribution of semantic threat intelligence.

IV. Model Optimization and Evaluation: The proposed framework is assessed using:

- Coefficient of determination (R^2)
- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)

Statistical dependability is ensured by using bootstrap resampling:

$$CI_{R^2} = Percentile_{[2.5,97.5]}(R^2(\mathcal{D}_b))$$

where, $\mathcal{D}_b$ define a bootstrap sample.

Furthermore, the ensemble significantly outperforms individual models, as indicated by paired statistical tests (t-test and Wilcoxon signed-rank test).

V. Interpretability and Continuous Improvement: To support TARA's continuous improvement objective:

- RF and XGB models are subjected to SHAP analysis in order to measure structured feature contributions.
- Influential threat phrases are identified by BERT using gradient-based token attribution.

Table 2: Hyperparameter Configuration of the Proposed Hybrid Multimodal Ensemble Model

Component	Hyperparameter	Configuration
Random Forest Regressor (RF)	Number of trees ($n_{\text{estimators}}$)	500
	Maximum depth	Unlimited
	Split criterion	Squared error
	Bootstrap sampling	Enabled
	Random state	42
	Parallel jobs	All available cores
XGBoost Regressor (XGB)	Number of boosting rounds ($n_{\text{estimators}}$)	500
	Maximum tree depth ((max_depth))	6
	Learning rate (η)	0.1
	Subsample ratio	0.8
	Column subsample ratio	0.8
	Objective function	Squared error
BERT Regression Model	Pretrained encoder	BERT-base
	Maximum sequence length	256
	Regression head	Linear (Dense, 1 unit)
	Loss function	Mean Squared Error
	Optimizer	AdamW
	Training epochs	4
	Batch size	8
	Gradient accumulation steps	2
	Mixed precision	Enabled (FP16)
LSTM Text Regression (Baseline)	Vocabulary size	30,000
	Embedding dimension	128
	LSTM hidden units	128
	Maximum sequence length	200
	Loss function	Mean Squared Error
	Optimizer	Adam
	Batch size	64
	Epochs	5
Ensemble Fusion	RF weight (α)	0.3
	XGB weight (β)	0.3
	BERT weight (γ)	0.4
	Fusion strategy	Weighted late fusion

IV. RESULTS AND DISCUSSIONS

A. Experimental Setup

All experiments were conducted using a reproducible and widely supported software environment suitable for cybersecurity data analysis. The entire workflow was implemented in Python using Jupyter Notebook, which allowed step-by-step experimentation and result tracking. Data preprocessing, cleaning, and feature extraction were performed using Pandas and NumPy libraries. Traditional machine learning experiments were implemented using the scikit-learn framework, while ensemble learning was supported through XGBoost and LightGBM libraries. For deep learning and transformer-based experiments, PyTorch was used along with the Hugging Face Transformers library. Visualization of results, including

performance comparison and statistical analysis, was carried out using Matplotlib to ensure clear and interpretable graphical outputs. The experiments were executed on a computing system equipped with a multi-core processor and adequate system memory to process a large CVE dataset efficiently. GPU acceleration was utilized for training transformer-based models to reduce computational time and support batch processing of textual data. The hardware configuration enabled stable execution of both traditional machine learning and deep learning tasks, ensuring consistent performance evaluation across all experimental stages.

B. Evaluation Metrics

Various regression-based evaluation metrics were employed to assess the performance of the proposed models, including R^2 score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). These metrics were selected because the objective of this work is to predict CVSS scores as continuous values. R^2 Score: R^2 indicates how well the predicted CVSS scores fit the actual values. A higher R^2 score represents better prediction performance.

Mean Absolute Error (MAE): MAE measures the average absolute difference between the predicted and actual CVSS scores. Lower MAE values indicate higher prediction accuracy.

$$MAE = (1 / n) \Sigma |y - \hat{y}|$$

C. Performance of the Models

Among all evaluated models, the proposed ensemble model (RF + XGB + BERT) demonstrated the best and most consistent performance across all evaluation metrics. As shown in Table 3, the ensemble achieved the highest R^2 value of 0.9947 along with the lowest MAE of 0.0133, which clearly indicates superior prediction accuracy and lower error compared to all other models. These results show that the ensemble model is able to learn both structured vulnerability characteristics and semantic information from textual descriptions effectively. Based on this strong and consistent performance, the ensemble model was selected as the final model for this research. Following the results presented in Table 3, it can be observed that the Random Forest model performs reasonably well, achieving a high R^2 value. This indicates that Random Forest is effective in capturing non-linear relationships among structured CVE features. However, its MAE is noticeably higher than that of the ensemble model, suggesting that prediction errors are still present when the model is used independently. Similarly, the XGBoost model shows good predictive capability, but its higher MAE compared to Random Forest and the ensemble indicates reduced robustness. Although XGBoost benefits from its boosting mechanism, it does not achieve the same balance between accuracy and error minimization as the ensemble model.

Table 3: Performance comparison of individual and ensemble models for CVSS score prediction

Model	R^2	MAE
Random Forest (RF)	0.994274	0.030703
XGBoost (XGB)	0.992037	0.049317
BERT Only	0.698428	0.648635
RF + XGB	0.994274	0.030703
Proposed Ensemble	0.994681	0.013250

The BERT-only model shows significantly weaker performance, with a much lower R^2 value and a very high MAE. This clearly indicates underfitting. Since this model relies only on textual vulnerability descriptions and does not use structured CVSS-related features, it fails to predict CVSS scores accurately. This result demonstrates that textual information alone is not sufficient for reliable risk estimation. The RF + XGB combined model improves performance compared to individual tree-based models; however, it still does not outperform the full ensemble. This suggests that partial model combinations are less effective than integrating both structured and textual representations together.

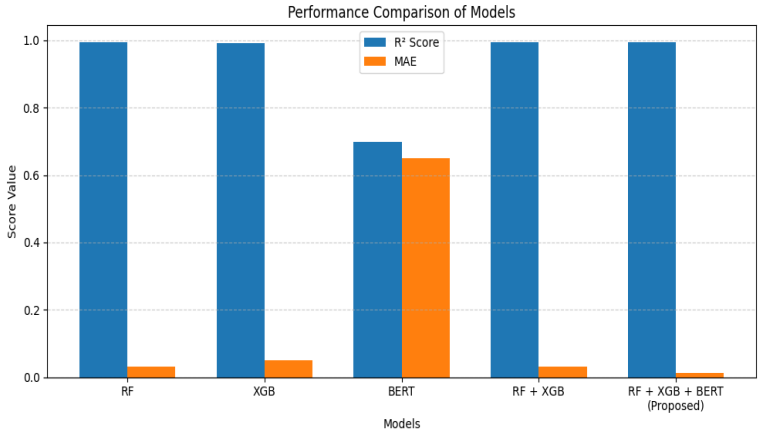

Fig. 3: Performance comparison of individual and ensemble models based on R^2 score and MAE.

Figure 3 presents a visual comparison of the performance of all evaluated models. The figure clearly shows that the proposed ensemble model consistently outperforms all other approaches in terms of accuracy and error reduction. Therefore, considering both quantitative results and comparative analysis, the RF + XGB + BERT ensemble model is chosen as the final and most reliable model for continuous CVSS-based cybersecurity risk assessment.

D. Error Distribution and Severity-Level Analysis

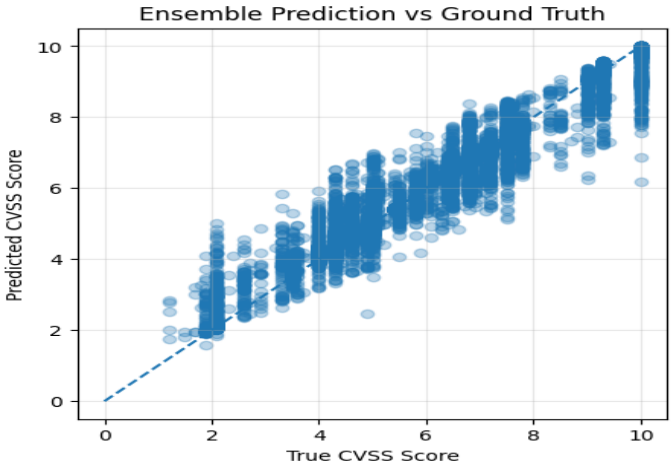

Fig. 4 : Ensemble Prediction vs Ground Truth

Figure 4 illustrates the relationship between the predicted CVSS scores produced by the proposed ensemble model and the corresponding ground truth values. The majority of data points are closely aligned

along the diagonal reference line, indicating a strong agreement between predicted and actual CVSS scores across the full severity range. This alignment confirms the high R^2 value achieved by the ensemble model and demonstrates its ability to accurately model both low- and high-severity vulnerabilities. A slightly wider dispersion can be observed in the higher CVSS ranges (above 8.0), which is expected due to the inherent complexity and diversity of critical vulnerabilities. However, no systematic deviation or prediction collapse is observed, indicating that the model does not suffer from bias toward overestimation or underestimation. This behavior is particularly important from a cybersecurity risk assessment perspective, as reliable estimation of high-severity vulnerabilities is crucial for prioritization and mitigation.

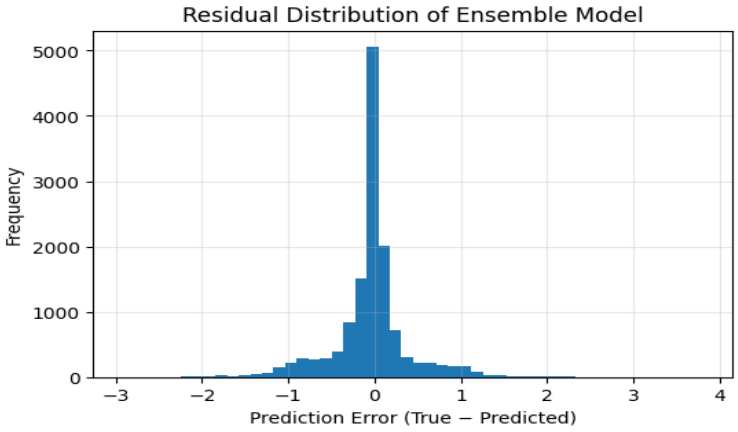

Fig. 5: Residual Distribution of the Ensemble Model

Figure 5 presents the residual error distribution of the proposed ensemble model, where the residual is defined as the difference between the true CVSS score and the predicted value. The residuals are symmetrically distributed around zero, with a sharp central peak, indicating that most predictions exhibit minimal error. This distribution suggests that the ensemble model achieves stable and unbiased predictions across the dataset. The absence of heavy skewness in the residual distribution further confirms that the model does not consistently over-predict or under-predict vulnerability severity. Although a small number of larger residuals are present in the tails, these correspond primarily to edge cases involving highly complex vulnerabilities. Overall, the narrow concentration of residuals around zero supports the robustness and generalization capability of the proposed ensemble approach.

E. Explainability Analysis of the Proposed Ensemble Model

Explainability is a critical requirement for deploying machine learning models in cybersecurity risk assessment, as analysts must understand the factors influencing automated severity predictions. To ensure transparency and analyst trust, the proposed ensemble model was analyzed using explainable AI techniques that separately examine the contribution of structured vulnerability attributes and textual semantic features. This dual-level analysis aligns with the principles of Threat Analysis and Risk Assessment (TARA).

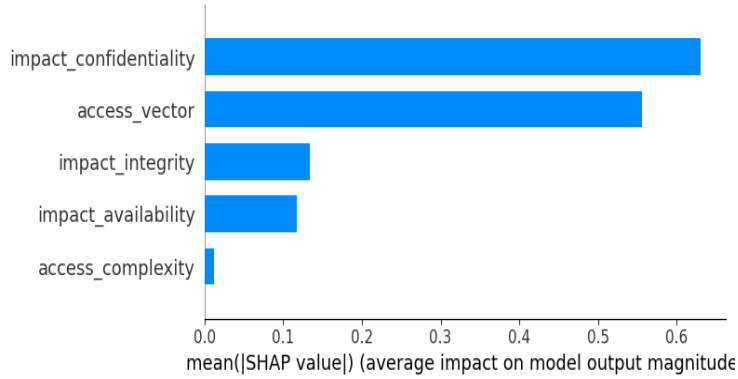

Fig. 6 : SHAP-based feature importance of structured vulnerability attributes

Figure 6 presents the SHAP-based feature importance analysis for the structured component of the proposed ensemble model, incorporating Random Forest and XGBoost predictors. The results indicate that *impact_confidentiality* and *access_vector* are the most influential features in determining the predicted CVSS score. This observation is consistent with CVSS design principles, where the potential impact on confidentiality and the mode of access play a dominant role in assessing vulnerability severity. In addition, *impact_integrity* and *impact_availability* contribute moderately to the prediction outcome, reflecting their role in measuring system compromise and service disruption. In contrast, *access_complexity* exhibits relatively lower influence, suggesting that while exploit difficulty is relevant, it is secondary to direct impact-related attributes. Overall, the structured feature analysis confirms that the ensemble model captures meaningful threat and impact characteristics that are central to cybersecurity risk assessment.

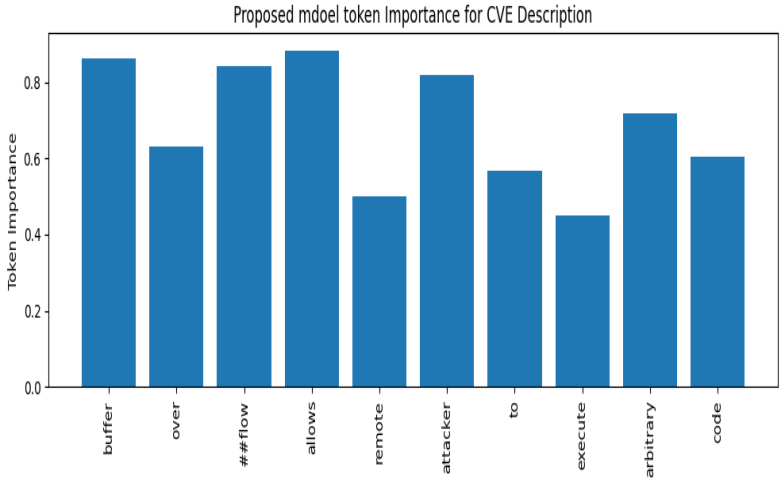

Fig. 7: Token-level importance visualization for CVE description

Figure 7 illustrates the token-level importance analysis of the BERT-based textual component of the proposed ensemble model. The visualization highlights that security-relevant terms such as buffer, overflow, remote, attacker, arbitrary, and code exert strong influence on the predicted CVSS scores. These tokens are directly associated with common exploitation techniques and high-impact attack scenarios, such as remote code execution. The presence of subword tokens further demonstrates BERT's capability

to capture fine-grained semantic representations from vulnerability descriptions. Importantly, the alignment between influential tokens and well-established cybersecurity terminology validates the model's ability to extract meaningful semantic cues rather than relying on spurious correlations. This behavior enhances analyst confidence in the textual reasoning process of the proposed ensemble.

V. CONCLUSION AND FUTURE WORK

This paper presents a continuous improvement framework for Threat Analysis and Risk Assessment (TARA) that incorporates Cybersecurity Threat Intelligence (CTI) and an intelligent, data-driven mechanism for predicting risks. Unlike conventional TARA approaches that rely on periodic risk evaluations, the proposed framework offers adaptive risk recalibration through continuous integration of new threat intelligence. A hybrid multimodal ensemble model was created to utilize the strength of the combined Random Forest, XGBoost, and BERT algorithms to effectively exploit the attributes of vulnerability data in both structured and unstructured threat descriptions for accurate CVSS score prediction. Experimental evaluation on a large-scale CVE dataset showed that the ensemble model performs significantly better than individual and combined models, with an R^2 value of 0.9947 and a Mean Absolute Error of 0.0133. The results show that the risk estimation mechanism, where semantic threat intelligence is integrated with structured vulnerability attributes, is more accurate and reliable.

In addition, the explainability analysis performed on the ensemble model using the SHAP and token-level attribution methods revealed that the risk estimation mechanism relies on important security-related features and threat semantics, thus ensuring its transparency. The proposed CTI-driven, continuous TARA model bridges the gap between the generation and risk assessment processes, thus making TARA an intelligent, resilient, and adaptive risk management model in the field of cybersecurity. This model can be extremely useful in vulnerability prioritization and decision-making in the presence of complex and dynamic threats. As future work, this framework can be extended by integrating real-time CTI feeds and attack graph modeling to identify the propagation paths of attacks. Moreover, the scalability and privacy aspects of the framework can also be improved by studying the online/federated learning mechanisms.

REFERENCES

1. Conti, M., Dehghantanha, A., Franke, K., & Watson, S. (2018). Internet of Things security and forensics: Challenges and opportunities. *Future Generation Computer Systems, 78*, 544–546.
2. Tounsi, S., & Rais, H. (2018). A survey on technical threat intelligence in the age of sophisticated cyber attacks. *Computers & Security, 72*, 212–233.
3. International Organization for Standardization. (2021). *ISO/SAE 21434: Road vehicles—Cybersecurity engineering*. ISO.
4. Shostack, A. (2014). *Threat modeling: Designing for security*. Wiley.
5. Mavroeidis, R., & Bromander, S. (2017). Cyber threat intelligence model: An evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence. In *Proceedings of the European Intelligence and Security Informatics Conference (EISIC)* (pp. 91–98). IEEE.
6. European Union Agency for Cybersecurity. (2022). *Cyber threat intelligence: ENISA good practice guide*. ENISA.
7. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
8. Husák, M., Komárková, J., Bou-Harb, E., & Čeleda, P. (2019). Survey of attack projection, prediction, and forecasting in cybersecurity. *IEEE Communications Surveys & Tutorials, 21*(1), 640–660.
9. Rahmati, M. (2025). *Towards explainable and lightweight AI for real-time cyber threat hunting in edge networks* (arXiv preprint). arXiv.

10. Sorokoletova, O., Antonioni, E., & Colò, G. (2025). *A scalable AI-driven framework for cyber threat intelligence information extraction* (arXiv preprint). arXiv.
11. Al-Yasiri, J. H., et al. (2025). *A threat intelligence event extraction conceptual model for CTI feeds* (arXiv preprint). arXiv.
12. Barakat, A. (2025). AI-driven threat intelligence for strengthening cyber defense mechanisms. *International Journal of Scientific Research Archives*.
13. Balasubramanian, P., et al. (2025). Generative AI for cyber threat intelligence: Applications and challenges. *Artificial Intelligence Review*.
14. Kwentoa, I. K. (2025). AI-driven threat intelligence for enterprise cybersecurity. *Journal of Next-Generation Research*.
15. Panda, S., et al. (2025). AI-driven predictive cyber threat intelligence framework. *Journal of Emerging Technologies and Innovative Research (JETIR)*.
16. Studies on AI-based cyber attack forecasting models. (2025).
17. Research on NLP-enhanced CTI automation pipelines. (2025).
18. Systematic review on cyber threat intelligence technologies and effectiveness. (2025). *Sensors*.
19. Fatima, N., Noorain, A., Ahmed, S. T., & Siddiqha, S. A. (2025, December). Automated Medical System for Rural Communities to Provide Medication without Human Interruption Using Machine Learning Techniques. In 2025 IEEE 5th International Conference on ICT in Business Industry & Government (ICTBIG) (pp. 1-5). IEEE.