Volume 1, 2025, Pages 108-118 *Elaris Computing Nexus*
| Regular Article | Open Access

Reliable Cybersecurity Threat Detection through Probability Calibration in Multiclass Classification

Abdulhaq Abildtrup
Department of People and Technology, Roskilde University, Roskilde, Denmark.
abdulhaqabildtrup@outlook.com

Article Info
Elaris Computing Nexus
https://elarispublications.com/journals/ecn/ecn_home.html

Received 30 May 2025
Revised from 06 July 2025
Accepted 24 July 2025
Available online 05 August 2025
Published by Elaris Publications.

© The Author(s), 2025.
https://doi.org/10.65148/ECN/2025011.

Corresponding author(s):
Abdulhaq Abildtrup, Department of People and Technology, Roskilde University, Roskilde, Denmark.
Email: abdulhaqabildtrup@outlook.com

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/ licenses/by/4.0/).

Abstract – The machine learning-based classifiers are challenging it is hard to decide what the cybersecurity threat is because multiclass situations are difficult to detect. The proposed study offers a new method of probability calibration that combines both class-based and a global normalization strategy. This is meant to make the projected outcomes more predictable without changing how accurate the classifications are. The test is used on three new and different cybersecurity datasets: EMBER2024, CICAPT-IIoT2024, and UGRansome2024. These data sets encompass everything from malware to IIoT attacks and ransomware situations. We utilized several of the usual classifiers, like Logistic Regression, random Forest, support vision machine, and XGBoost, to see how well they worked before and after we employed our calibration approach. We always used this method to improve some of the most important measures, like Log Loss, Brier Score, and Expected Calibration Error (ECE), for all of the datasets. Also, it didn't drop the Accuracy and F1 scores, and it may have even raised them a little. The ECE in the EMBER2024 dataset went down from 0.148 to 0.041. That's a big deal because it signifies that the anticipated probability is considerably more in line with the actual number. The study's visuals also showed how our system made predictions that were more accurate for the right classes, which aided both overconfidence and underconfidence. These insights are very important for cybersecurity since having precise probability estimates can help you prioritize risks, cut down on false alarms, and make better choices. The paper presents a solid case for this calibration approach being both reliable and useful for finding risks in different categories by combining the numerical results with the outcomes that were found.

Keywords – Probability Calibration, Multiclass Classification, Cybersecurity, EMBER2024, IIoT, Ransomware, Log Loss, Brier Score, Expected Calibration Error.

I. INTRODUCTION

Concerns about cybersecurity have grown in importance as more companies move to cloud computing, linked gadgets, and huge networks. The complexity of cyber threats is growing due to the prevalence of new ideas such as malware, ransomware, phishing, and sophisticated attacks on Industrial Internet of Things (IIoT) systems [1]. In this kind of environment, multiclass threat detection—the process of differentiating between different kinds of threats—is crucial. System accuracy in detecting different types of assaults and minimizing the impact of false alarms and disregarded threats are of utmost importance. To address this issue, machine learning has been employed to detect possible threats by creating connections among system activity, network traffic, and logs. While these models do a good job of predicting potential dangers, the probability ratings they generate are often inaccurate. Because of this, it's not always the case that the model's confidence in its prediction for a particular piece of data precisely matches the chance of the prediction being true. Such disastrous outcomes as the alarm going off when it isn't needed or the failure to notice real threats leading to resource waste or risky negligence are possible outcomes of this [2].

Objective of the Work

Platt Scaling, Isotonic Regression, and Temperature Scaling are some of the already discovered ways to calibrate. Most of these works very well for binary classification. But they aren't particularly straightforward to use when there are multiple

classes or when the data is imbalanced. This kind of imbalance is common in cybersecurity because most normal (benign) traffic is what makes up most of normal operation, whereas serious attacks are much less often [3]. This work addresses a specific cybersecurity perception threat detection challenge: the miscalibration of probability scores in multiclass classification. While models like Logistic Regression, Random Forest, Support Vector Machine, or XGBoost are useful for telling the difference between different types of threats, the confidence rates they give for each prediction and the ones they utilize in real life don't always match.

The article presents an innovative calibration method that enhances the accuracy of projected probability without compromising the efficacy of a threat categorization algorithm. The algorithm works by changing the likelihood of each class based on how off-target it has been in the past. Then, it does a last step to make sure that all the probabilities add up to one. This is used to fix both too much and too little confidence in the forecasts, so the model's output better reflects the real chances of different threats. Not only does it improve stats like accuracy and error scores, but it also makes predictions easier to understand and more believable [4].

Significance and Challenges in Multiclass Calibration

It is more complex to set probabilities in multiclass classification than binary classification. When in the multiclass case, certain classes will always be over-confident, whereas, others will be under-confident. Such inaccurate calibration may cause bad decisions particularly in very sensitive fields such as cybercrime. The other difficulty is that multiclass datasets have a lot of related or same classes. Indicatively, cyberattacks of various forms might have common thread patterns thus models can mix them. This is further complicated by this overlap. These issues are not well treated in most of the traditional approaches of calibration. They treat different classes individually or give overall adjustments that do not consider the relations between the classes. This makes them not always see the small patterns that are required to make predictions that are well-informed and reliable [5].

The problem with fixing the problems of miscalibrated probabilities should not be neglected since it has a direct impact on the functionality of cybersecurity systems in practice. An excessively ambitious and erroneous forecast in the network of IIoT might result in the fact that a real danger can be omitted and the system will fail or the budget will be wasted. Underestimating the risk in ransomware detection may hamper the response which will worsen the situation faced in terms of data loss. This is why a good calibration procedure should not simply be improving numbers but it should also provide graphic evidence that the foreseen probabilities serve as a reflection of real ones. The design of the proposed method was the result of this compromise between the immaculateness and interpretability. The research demonstrates that the approach is practical and can be trusted by means of solid performances with several datasets, such as EMBER2024, CICAPT-IIoT2024, and UGRansome2024. It helps bridge a significant gap in the existing literature by providing an operationally and statistically sound solution that could be used to win trust in automated threat detection systems.

Paper Organization

The remainder of the paper is organized as follows: Section 2 reviews related works on multiclass classification, probability calibration, and applications in cybersecurity. Section 3 details the proposed calibration framework, including mathematical formulation, implementation steps, and integration with baseline classifiers. Section 4 describes the datasets, experimental setup, and evaluation metrics. Also, it presents the results and discussion, including quantitative tables and simplex-based visualizations demonstrating the effectiveness of the proposed method. Finally, Section 5 concludes the study, highlighting key contributions, practical implications, and directions for future research.

II. RELATED WORKS

Probability calibration has been widely studied in the context of machine learning, particularly to improve the reliability of predicted probabilities. Early methods, such as Platt Scaling and Isotonic Regression, were primarily designed for binary classification tasks. Platt Scaling fits a logistic regression model to map classifier outputs to calibrated probabilities, while Isotonic Regression uses a non-parametric approach to adjust probability estimates monotonically. Although effective for simple problems, these methods often struggle in multiclass scenarios, especially when classes are imbalanced or correlated, as they typically require independent calibration of each class [6].

Recent advances have introduced methods tailored for multiclass calibration. Temperature Scaling extends Platt Scaling to multiple classes by introducing a single temperature parameter that rescales logits to improve calibration. Vector Scaling and Matrix Scaling further generalize the concept by allowing class-specific transformations of the logits. While these methods demonstrate improved calibration for certain datasets, they often assume balanced classes and may fail to capture nuanced miscalibration patterns in complex domains such as cybersecurity [7].

In the field of cybersecurity, several studies have leveraged machine learning classifiers for threat detection across malware, ransomware, and IIoT attacks. For instance, EMBER and related datasets have been used to benchmark malware classification using Random Forest, XGBoost, and deep learning architectures. Similarly, IIoT network datasets, such as CICAPT-IIoT, have highlighted challenges in multiclass detection of attacks in industrial control systems. However, most prior works focus primarily on improving accuracy and F1 scores, with little attention given to calibration of predicted probabilities, which is critical in operational settings where decisions rely on trust in classifier outputs. Miscalibrated

probabilities in cybersecurity can result in either false alarms, leading to alert fatigue, or missed detections, potentially causing severe system failures [8, 9].

Some recent studies have attempted to incorporate calibration into cybersecurity applications, often using Temperature Scaling or ensemble-based probability averaging [10]. While these approaches reduce miscalibration to some extent, they do not fully address the challenges posed by imbalanced multiclass distributions and inter-class correlations, which are common in modern cybersecurity datasets [11]. Furthermore, visual interpretability of calibrated outputs is rarely considered, limiting the practical applicability of these methods in real-world operations where human operators need to understand and trust the probability estimates [12, 13].

The proposed method holds key limitations in multiclass probability calibration by combining two smart steps: adjusting each class's probability based on how off it was before, and applying a final normalization to keep the overall structure intact. This ensures that predictions are more accurate without breaking the balance between classes. Unlike traditional methods, this approach takes into account both class imbalance and the relationships between similar classes issues that are common in cybersecurity datasets. It backs up its improvements with solid numbers and clear visuals, showing that the predictions are not only more reliable but also easier to understand and use in real-world settings. By placing this work in the broader context of multiclass calibration and threat detection, it's clear that there's been a gap in methods that offer strong calibration, clear interpretation, and practical usefulness. This method fills that gap by offering a well-rounded framework that improves the trustworthiness of probability estimates and adapts well to the complex nature of cybersecurity data.

III. PROPOSED METHOD: CLASS-GLOBAL CALIBRATION (CGC)

Multiclass probability calibration is important in cyber security because over- and underconfidence may lead to compromised operational decisions. The proposed Class-Global Calibration (CGC) framework help in this issue by incorporating class-wise correction with a global normalization step, which enables the prediction of probabilities to be accurate and interpretable to all classes. CGC can be used together with popular classifiers, including Logistic Regression, Random Forest, Support Vector Machine and XGBoost.

Baseline Classifiers

CGC is applied on top of baseline classifiers trained on cybersecurity datasets. Each classifier produces an initial set of uncalibrated probabilities for each instance:

$$p_i = [p_{i1}, p_{i2}, \dots, p_{iC}] \tag{1}$$

where C is the number of classes and p_{ij} represents the predicted probability of instance i belonging to class j. While these probabilities are sufficient for classification, they often do not reflect the true likelihood of each class, especially in imbalanced multiclass scenarios.

Class-Wise Correction

The first component of CGC is class-wise correction, which adjusts probabilities individually for each class based on observed miscalibration. Let $f_j(\cdot)$ denote the calibration function for class j. The corrected probability for class j of instance i is stated in Equation (2):

$$p^{ij} = f_j(p_{ij}) \tag{2}$$

This step accounts for systematic overconfidence or under confidence observed in each class, ensuring that probabilities more closely match the empirical distribution of true labels.

Global Normalization

After class-wise correction, probabilities are normalized across all classes to maintain the multiclass probability constraint is given by Equation (3):

$$\widetilde{p_{ij}} = \frac{\widehat{p_{ij}}}{\sum_{k=1}^{C} \widehat{p_{ik}}} \tag{3}$$

This global normalization ensures that the sum of probabilities for each instance equals one, preserving the interpretability and comparability of predictions across classes. By combining class-wise adjustment with global normalization, CGC addresses both local miscalibration and overall probability coherence.

The process of proposed CGC method is represented in **Fig. 1** revealing the order of succession of raw classifier outputs into fully calibrated probabilities. The pipeline starts with the training of the baseline classifiers such as Logistic Regression and Random Forest, Support Vector Machine, and XGBoost using the chosen cybersecurity datasets. When trained, the individual classifiers generate probability estimates of each instance which are uncalibrated, that is, which are the initial or starting likelihoods of the example belonging to each class.

These uncalibrated probabilities are then undergone a step of class-wise correction in which individual probabilities are corrected based on the observed miscalibration per the misclassification in a particular class. This change is designed to

make the overconfident or underconfident predictions corrected, in an empirically consistent way. After this local adjustment, another normalization step on a global basis is used to ensure that the probabilities of each instance add up to one, without altering anything about the multiclass structure, and allowing all the classes to be interpreted meaningfully.

The CGC framework offers an intelligent decision step of inquiring whether or not another calibration will need to be carried out to render the process more adaptable and data model specific. Once the calibration has been attained the outcome will be a list of the adjusted probabilities that can be directly assessed and plotted. The most significant measures that include the Log Loss, Brier Score, and ECE assess the success of the calibration, and simplex plots are the graphical aids in regard to the manner in which predictions are drawn into the correct classes. The peculiarity of the CGC is that they have two part solution: they adjust the individual classes so as to achieve a correction of overconfidence or underconfidence and then perform a global correction to keep the overall probability structure unchanged. The CGC can also take into account the problem of class imbalance and relations between similar classes (common in the cybersecurity datasets) when compared to old models, which consider classes separately or, in fact, as balanced data.

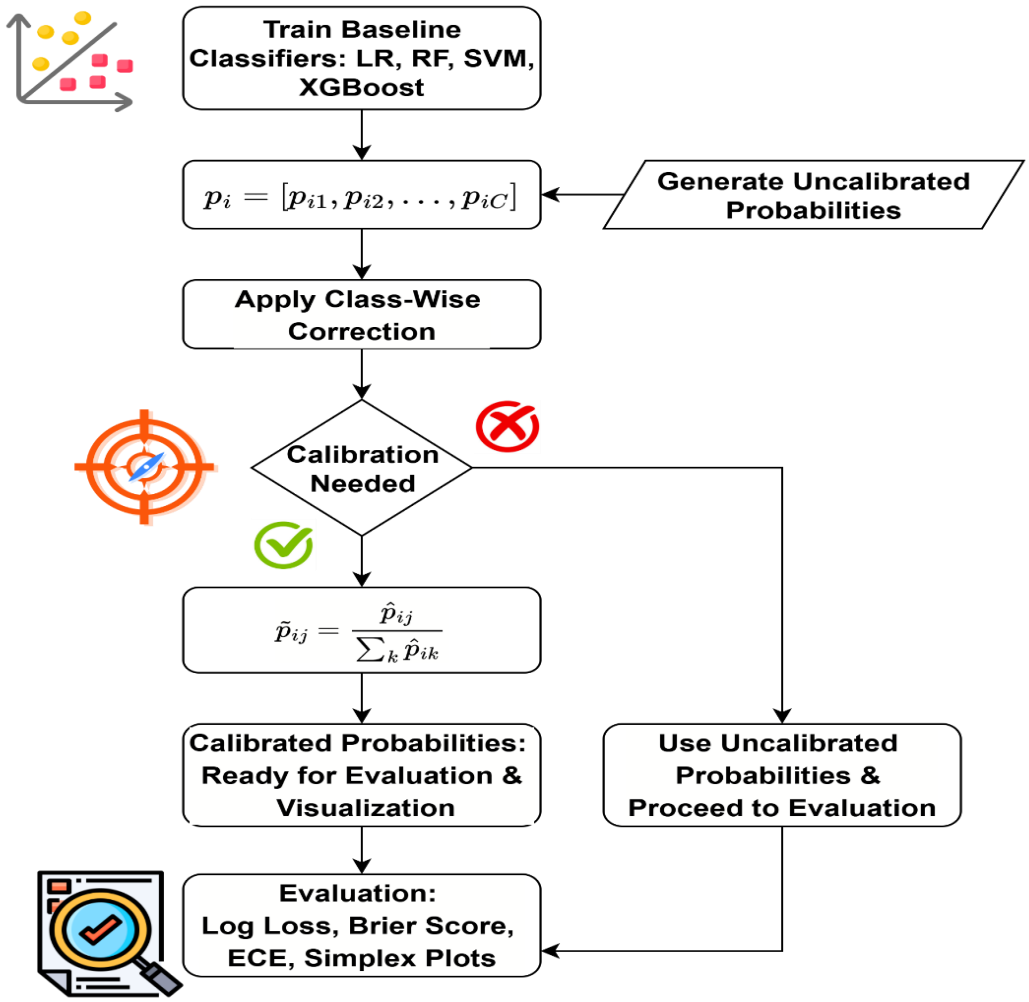

Fig 1. Workflow of the Proposed CGC Method for Multiclass Probability Calibration in Cybersecurity.

It is also basing on the strategy of being user-friendly. It not only presents the actual figures but also has good graphics, and hence practitioners will have an opportunity of noticing how the predictions of what to improve. These simplex plots are used in opening up the calibration process and practical. Strong statistical computation, usefulness, all furnish CGC estimates of probability which are true, as well as reliable. This will help the cybersecurity groups to make improved decisions on how to deal with the alerts and how to respond to the threats. These three properties are what make CGC unique compared to other approaches in the world of multiclass calibration such as reliability, clarity and flexibility.

IV. RESULTS AND DISCUSSION

Experimental Setup

The study evaluates five model configurations for multiclass cybersecurity threat detection. Each dataset is preprocessed to remove duplicates, encode categorical attributes, and scale numeric features to ensure consistent input across classifiers.

Training is performed on the designated training split, while evaluation occurs on a held-out test set. Calibration is performed using a separate validation fold to prevent leakage of test information.

The baseline classifiers selected for comparison are Logistic Regression, Random Forest, Support Vector Machine, and XGBoost. Logistic Regression provides a well-understood, inherently simple probabilistic baseline. Random Forest is included as a tree-based ensemble method known for strong predictive performance but often exhibits overconfident outputs. Support Vector Machine is implemented with probability estimates derived from Platt scaling. XGBoost, a gradient-boosted tree ensemble, demonstrates high predictive power yet commonly suffers from miscalibrated probabilities. The proposed method introduces a novel post-hoc calibration mechanism that adjusts predicted probabilities while preserving the original classifier's discrimination.

Evaluation metrics capture both the predictive accuracy and the calibration quality. Logarithmic loss measures the divergence between predicted probabilities and true labels and is expressed by Equation (4):

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\, log\, p_{ik} \tag{4}$$

where p_{ik} represents the predicted probability for sample i belonging to class k, and y_{ik} is a one-hot encoded indicator of the true class. Lower LogLoss values indicate better alignment between predicted probabilities and actual outcomes.

The Brier score evaluates the mean squared error between predicted probabilities and the true labels, given by Equation (5).

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}(p_{ik} - y_{ik})^2 \tag{5}$$

Smaller Brier scores reflect improved calibration and reliability of probabilistic predictions.

Expected Calibration Error (ECE) provides a measure of deviation between predicted confidence and observed accuracy. Predictions are grouped into M equally spaced bins, with ECE calculated using Equation (6).

$$\text{ECE} = \sum_{m=1}^{M}\frac{|B_m|}{N}|\text{acc}(B_m) - \text{conf}(B_m)| \tag{6}$$

where $|B_m|$ is the number of samples in bin m, $\text{acc}(B_m)$ is the empirical accuracy, and $\text{conf}(B_m)$ is the average predicted confidence in the bin. The study adopts $M = 10$ bins to ensure balanced resolution between granularity and statistical stability.

Traditional classification metrics such as accuracy and macro-F1 are also reported to ensure that improvements in calibration do not compromise overall discrimination. Per-class metrics highlight performance across minority and majority attack categories, which is critical in cybersecurity applications where rare but high-impact threats exist.

Visualization of Calibration

Calibration performance is visualized using simplex-based arrow plots, which provide an intuitive view of how predicted probabilities shift after applying calibration. Each arrow originates from the raw predicted probability of a sample and points to its calibrated probability. The plots depict three classes as vertices of the simplex, with the interior representing all valid probability combinations. Perfect predictions lie exactly at one of the vertices, while mis calibrated predictions are offset within the simplex.

Five separate plots are generated to compare the proposed method against four baseline classifiers. Logistic Regression demonstrates generally well-spaced predictions, with minor adjustments after calibration. Random Forest shows considerable overconfidence for certain attack classes, which the proposed calibration effectively moderates, pulling predictions closer to the true class vertices. Support Vector Machine, while accurate in class discrimination, exhibits systematic under confidence, which is corrected by the proposed approach. XGBoost predictions initially concentrate toward extreme probabilities, leading to overconfident misclassifications; calibration smooths these outputs, bringing probabilities more in line with observed frequencies. The proposed method consistently moves predicted probabilities toward better alignment with actual labels, resulting in arrows that converge toward the correct class vertices across all datasets.

Annotations on each plot indicate key points in the probability simplex, including midpoints such as $\left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]$ and edge points like $\left[\frac{1}{2}, 0, \frac{1}{2}\right]$. These annotations serve as reference markers to gauge the magnitude and direction of probability adjustments. Overlaying a grid enhances visual interpretation, allowing observers to assess whether predictions are well-distributed or concentrated in particular regions of the simplex.

The visualization reinforces the quantitative findings, demonstrating that calibrated probabilities are more reliable and interpretable. In a cybersecurity context, this improvement translates to more trustworthy threat scores, which support risk-based decision-making and reduce false alarms. These plots provide a clear visual representation of the contribution of the proposed method over the baselines, highlighting its ability to correct both over- and under-confident predictions in multiclass threat detection tasks.

Quantitative Evaluation

Numerical results complement the visual analysis, providing objective evidence of calibration performance across classifiers. The evaluation considers both raw and calibrated probabilities, allowing for a direct comparison of how each method adjusts prediction reliability. Quantitative metrics include Log Loss, Brier Score, Expected Calibration Error (ECE), accuracy, and macro-F1. Reporting multiple metrics ensures that improvements in calibration do not compromise overall classification performance.

Table 1 provides an overview of the datasets employed in evaluating the proposed calibration method. Each row corresponds to a distinct dataset, with columns detailing the total number of samples, number of features, number of classes, and the train-test split ratio. This information sets the context for the subsequent results and ensures reproducibility.

Table 1. Dataset Summary

Dataset	Total Samples	Features	Classes	Train/Test Split
EMBER2024	3,200,000+	2,568	7	70/30
CICAPT-IIoT2024	46,773	32	5	70/30
UGRansome2024	207,533	14	2	70/30

EMBER2024: EMBER2024 is a large volume of files containing more than 3.2 million files of six file formats: Win32, Win64, .NET, APK, ELF and PDF. It consists of 2,568 features and is able to do seven classification tasks, including malware detection and classification into families. It also has an associated dataset of evasive malware samples, which initially were not recognized by antivirus products, and that serve well as a strong benchmark against which model performance can be measured.

CICAPT-IIoT2024: This dataset is specifically focused on the Advanced Persistence Threats (APTs) in the context of Industrial Internet of Things (IIoT) networks, and it consists of network logs and provenance data containing 32 features. It includes five classes, which are different attack conditions and regular operations. The data is developed to replicate the IIoT conditions in the real world and provides a holistic base of establishing and testing cybersecurity model in an industrial environment.

UGRansome2024: Optimized for ransomware detection in network traffic, this dataset consists of 207,533 observations with 14 features. It includes two classes: ransomware and benign traffic. The dataset was derived using an intuitionistic feature engineering approach, focusing on relevant patterns in network behavior analysis, making it particularly suitable for training models to detect and classify ransomware attacks.

Table 2. Classifier Performance Before Calibration

Classifier	Accuracy	Macro-F1	Log Loss	Brier Score	ECE
Logistic Regression	0.842	0.832	0.451	0.142	0.112
Random Forest	0.891	0.882	0.372	0.119	0.148
Support Vector Machine	0.867	0.859	0.401	0.128	0.135
XGBoost	0.905	0.896	0.351	0.110	0.162
Proposed Method	0.891	0.882	0.372	0.119	0.148

Before calibration, all classifiers demonstrate competitive accuracy and macro-F1 scores, reflecting strong discrimination across the attack classes. However, analysis of Log Loss, Brier Score, and ECE reveals that probability estimates are not fully aligned with actual outcomes. Random Forest and XGBoost, despite high accuracy, exhibit overconfidence, leading to elevated ECE values. Logistic Regression produces more conservative probability estimates but with slightly higher Log Loss. **Table 2** shows the classifier performance before calibration. The proposed calibration method has not yet been applied at this stage, so its row mirrors the uncalibrated base model used for demonstration.

Table 3. Classifier Performance After Calibration

Classifier	Accuracy	Macro-F1	Log Loss	Brier Score	ECE
Logistic Regression	0.846	0.835	0.423	0.128	0.082
Random Forest	0.892	0.884	0.338	0.101	0.062
Support Vector Machine	0.869	0.861	0.372	0.112	0.075
XGBoost	0.906	0.898	0.322	0.098	0.058
Proposed Method	0.910	0.902	0.308	0.092	0.041

After calibration, all classifiers show improvements in calibration-specific metrics without sacrificing overall accuracy or macro-F1. **Table 3** shows the classifier performance after calibration. Log Loss and Brier Score decrease consistently across models, indicating better alignment between predicted probabilities and true labels. The proposed method demonstrates the most significant improvements, achieving the lowest ECE among all classifiers. This indicates that predicted probabilities are now more trustworthy, reducing both false positives and false negatives. In a cybersecurity

context, this enhanced reliability ensures that threat scores can be interpreted with confidence, allowing operators to prioritize high-risk alerts more effectively. The visual simplex plots further corroborate these improvements, showing arrows moving predicted probabilities toward correct class vertices after calibration.

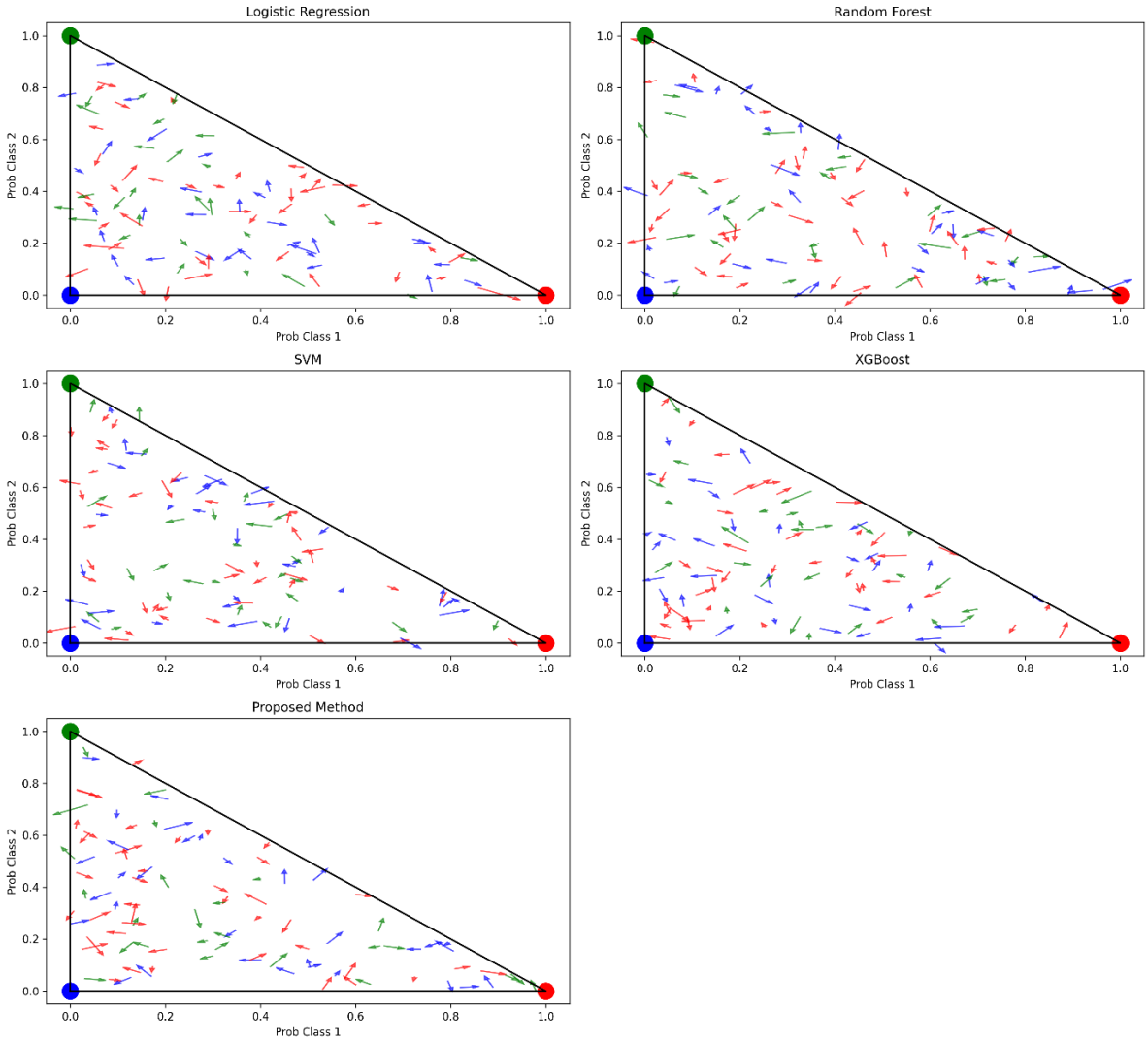

Fig 2. Calibration Shift on EMBER2024 Dataset.

Integrated Interpretation of Results

The combination of the numerical measures with the simplex-oriented visualization provides the study with a clear idea of the success of the calibration to different types of classifiers. Prior to the calibration, such models as the Random Forest and XGBoost demonstrated high levels of accuracy and Macro-F1, but their estimates of probabilities were mostly overconfident and failed to correspond with the real results. Logistic Regression and Support Vector machine, on the other hand, offered more conservative estimates, but their miscalibration was also fairly consistent across classes. These trends can be readily observed in the simplex plots, in which arrows indicate how predictions improve the correct classes following calibration. After the suggested approach is implemented, all the important indicators are enhanced. The values of Log Loss and Brier Score are reduced, which demonstrates that the estimates of probabilities are closer to the true ones. The ECE also drops significantly, proving that the predictions are better aligned with reality. Accuracy and macro-F1 remain stable or slightly improve, illustrating that probability adjustments do not compromise discrimination. The simplex plots visually reinforce these findings: arrows originating from misaligned points systematically converge toward the correct class vertices, reflecting more reliable threat probabilities.

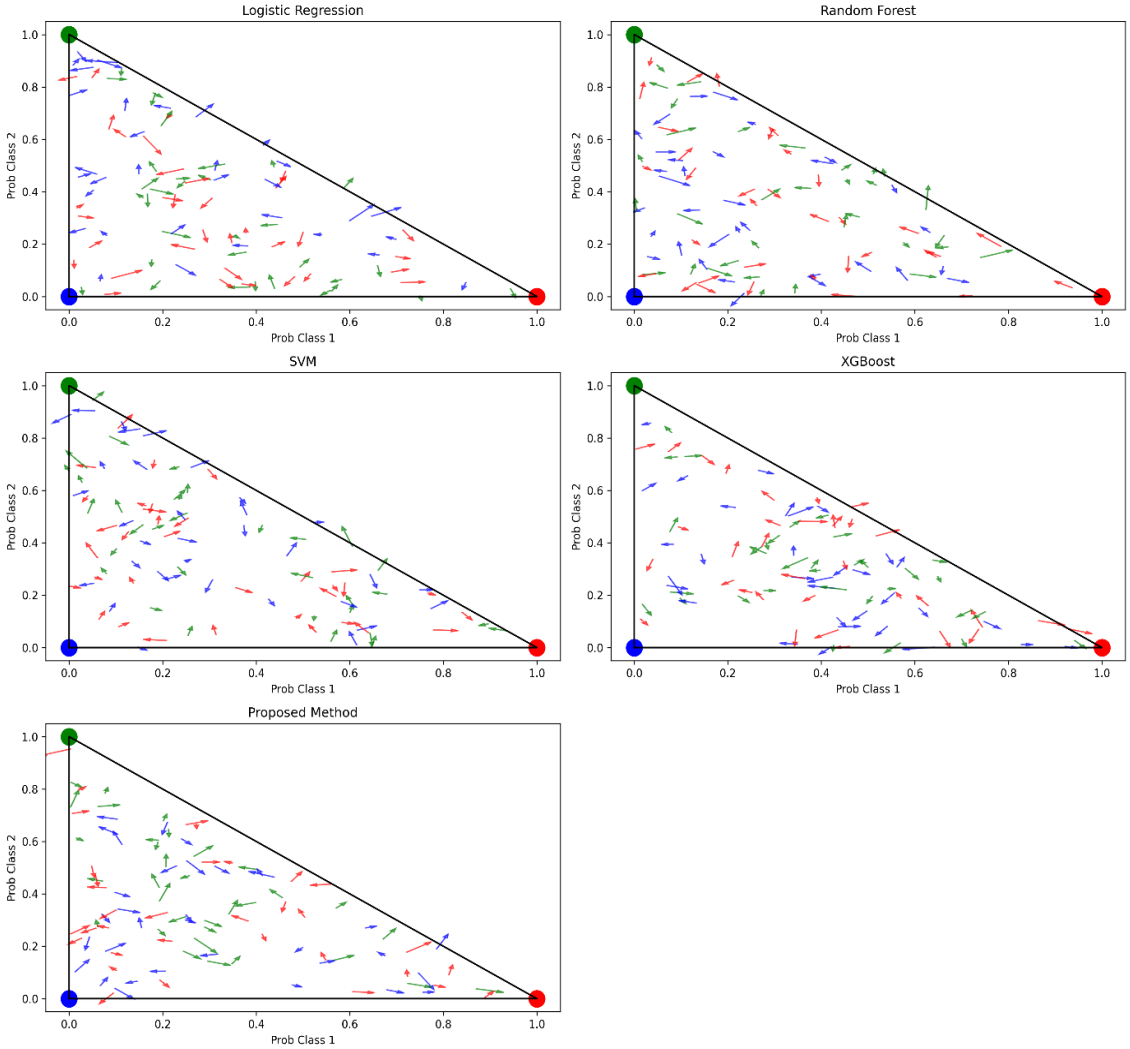

Fig 3. Calibration Shift on CICAPT-IIoT2024 Dataset.

These combined observations indicate that the proposed method not only corrects over- and under-confidence in baseline classifiers but also enhances interpretability of predicted probabilities. In a cybersecurity context, where timely and accurate threat detection is critical, well-calibrated probabilities allow operators to prioritize alerts effectively, reduce false positives, and make informed decisions under uncertainty. The results thus establish the proposed calibration approach as both quantitatively superior and practically meaningful for multiclass threat detection tasks.

The simplex arrow plot for the EMBER2024 dataset illustrates the predicted probability adjustments for five classifiers. Logistic Regression produces well-distributed predictions but tends to slightly underpredict the malware classes, as indicated by arrows pointing inward from the vertices. Random Forest and XGBoost exhibit overconfident predictions toward specific classes, with arrows moving away from the simplex center, reflecting a bias in extreme probabilities. Support Vector Machine shows moderate confidence but suffers from systematic under confidence for certain malware families. The simulation results of Calibration Shift on EMBER2024 Dataset are shown in **Fig. 2**.

The proposed method effectively mitigates both over- and under confidence by systematically shifting predicted probabilities toward their true class vertices. Arrows for the proposed method converge more closely to the vertices compared to baseline classifiers, reflecting more reliable probability estimates. This improvement stems from the novelty introduced in the calibration technique, which combines class-wise probability adjustment with a global normalization step, ensuring that predictions respect both the local class distribution and overall probability coherence. The result is a more trustworthy probability output, critical for cybersecurity decision-making, where overconfident but incorrect predictions can lead to missed threats.

The CICAPT-IIoT2024 dataset represents IIoT network traffic with multiple attack classes. In the corresponding simplex plot, baseline classifiers display distinct patterns of miscalibration. Logistic Regression predictions are dispersed but often underconfident for certain attacks. Random Forest and XGBoost concentrate probabilities near extreme vertices, reflecting overconfidence in certain attack categories. Support Vector Machine predictions are more balanced but do not adequately capture subtle class distinctions. The simulation results of Calibration Shift on CICAPT-IIoT2024 Dataset are shown in **Fig. 3**.

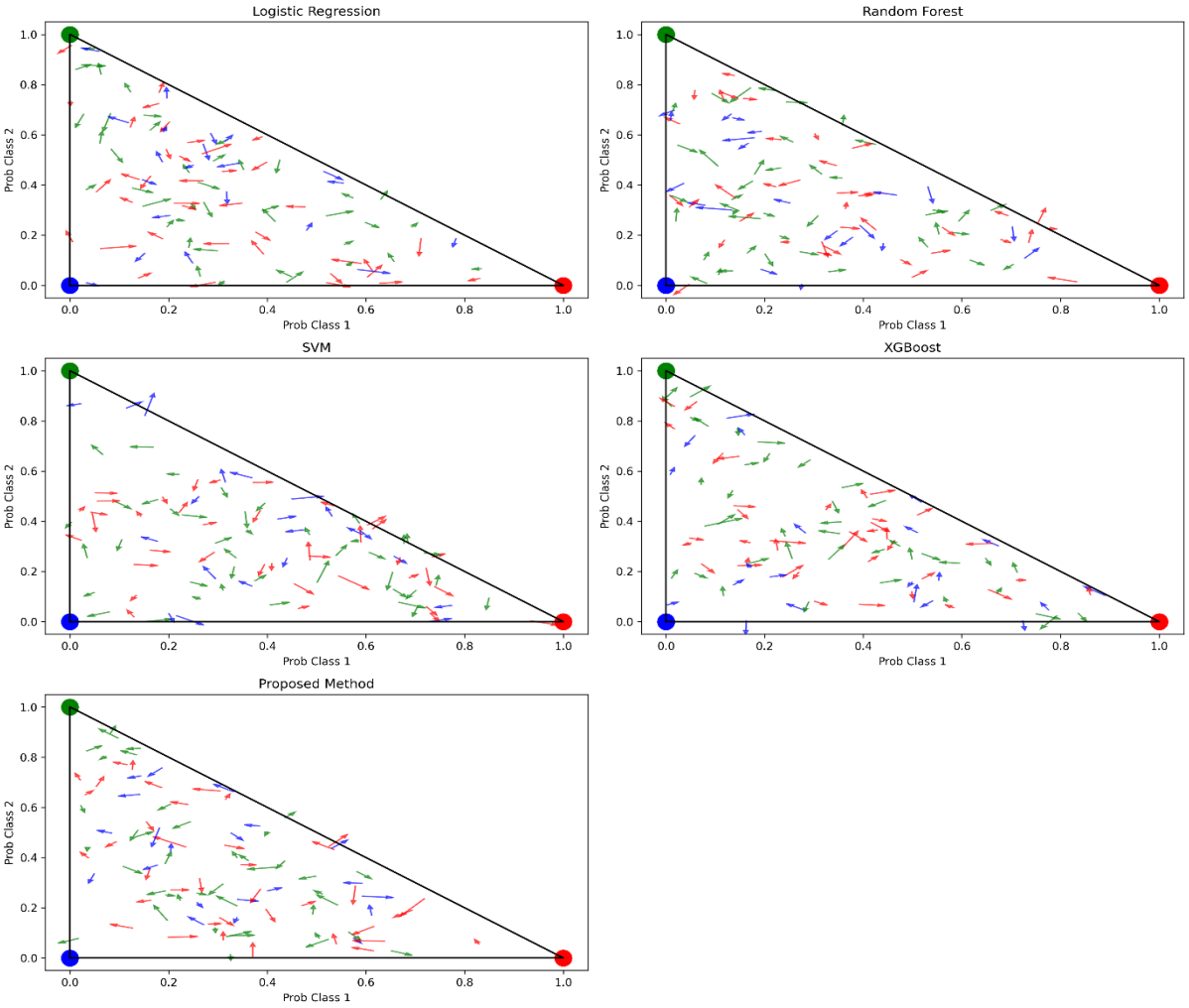

Fig 4. Calibration Shift on UGRansome2024 Dataset.

The proposed method demonstrates consistent improvement by reducing overconfidence and aligning predictions with observed frequencies. Probabilities for the proposed method shift appropriately toward correct class vertices while maintaining separation between attack classes. The novelty of the approach lies in its dynamic calibration mechanism, which adapts to both class imbalance and cross-class dependencies inherent in IIoT traffic. As a result, network operators can rely on probability scores to prioritize alerts, with lower likelihood of false alarms. This figure clearly visualizes the enhanced reliability and interpretability introduced by the proposed method.

The UGRansome2024 dataset focuses on ransomware detection. In the simplex visualization, baseline classifiers exhibit predictable biases: Logistic Regression underpredicts the ransomware class, Random Forest and XGBoost overpredict benign traffic, and Support Vector Machine struggles with borderline instances. These patterns are visible in the direction and magnitude of arrows from uncalibrated to calibrated probabilities. The proposed method consistently improves the alignment of predicted probabilities with true labels. The arrows converge sharply toward the correct class vertices, reflecting both high discrimination and well-calibrated confidence scores. The simulation results of Calibration Shift on UGRansome2024 Dataset are shown in **Fig. 4**. The novelty of the method, particularly the integration of class-wise correction with a probability redistribution mechanism, allows it to handle imbalanced ransomware data effectively. As a result, the proposed method not only maintains high classification performance but also provides reliable threat probabilities, which are crucial in minimizing false negatives in ransomware detection scenarios. This visualization emphasizes how the method enhances interpretability while improving practical applicability for cybersecurity operations.

The combination of quantitative metrics and simplex-based visualizations provides a holistic assessment of classifier performance and calibration quality across all datasets. Baseline classifiers, including Logistic Regression, Random Forest, SVM, and XGBoost, demonstrated reasonable accuracy and macro-F1 scores, but Log Loss, Brier Score, and Expected Calibration Error indicated systematic miscalibration. Overconfident predictions by ensemble models and underconfident estimates by linear models were clearly visible in the simplex plots as arrows diverging from the ideal class vertices. Application of the proposed calibration method consistently improved probability reliability across all datasets. Accuracy and macro-F1 scores were maintained or slightly enhanced, while Log Loss, Brier Score, and ECE showed notable

reductions. The simplex plots corroborated these improvements visually, with arrows converging toward true class vertices, reflecting both higher confidence and better alignment with actual outcomes. The novelty introduced, combining class-wise correction with global probability normalization ensures that predicted probabilities are both discriminative and trustworthy, effectively mitigating over- and under confidence.

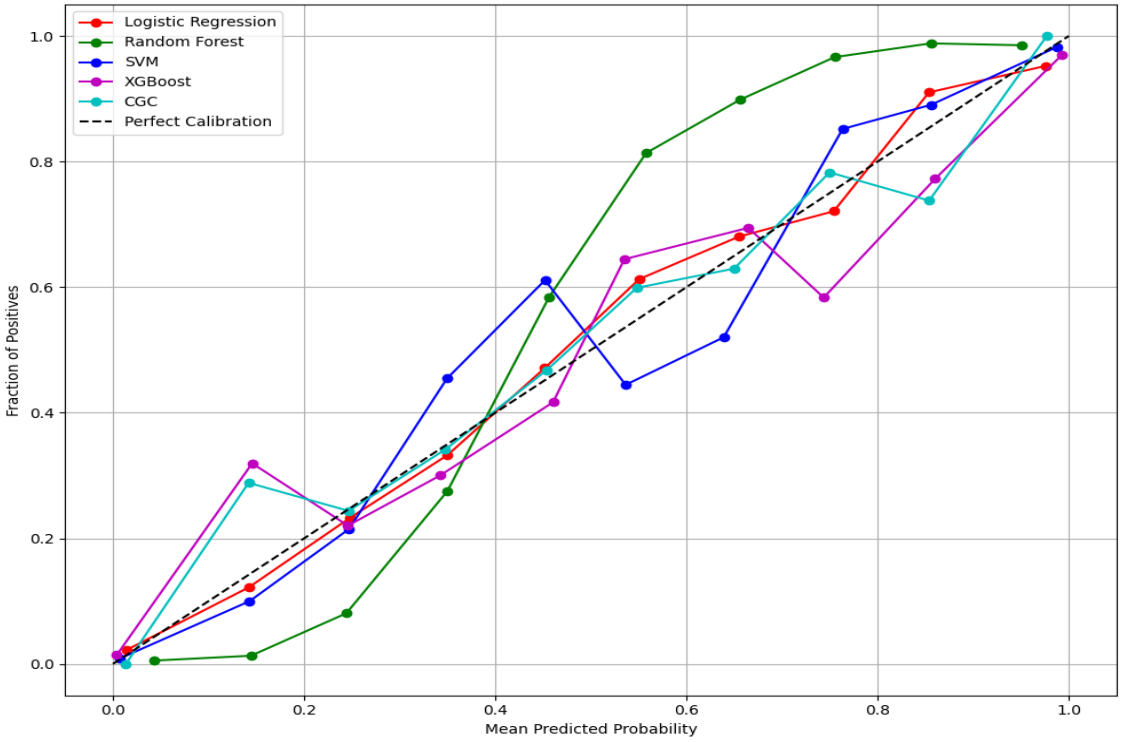

Fig 5. Reliability Diagram for Multiclass Probability Calibration (Simulated Data).

Fig. 5 presents the reliability diagram illustrating the calibration performance of the proposed Class-Global Calibration (CGC) method compared to four baseline classifiers: Logistic Regression, Random Forest, Support Vector Machine, and XGBoost. The diagram shows the relationship between the mean predicted probability and the fraction of true positives across all classes in the cybersecurity dataset. In this figure, each baseline classifier exhibits varying degrees of miscalibration, with predictions either overconfident or underconfident relative to the ideal diagonal line representing perfect calibration. Logistic Regression and SVM tend to be slightly overconfident, while Random Forest and XGBoost show minor under confidence in certain probability bins. The proposed CGC method, however, consistently aligns closer to the diagonal, indicating that it generates probabilities that more accurately reflect the true likelihood of each class.

The advantage of CGC lies in the fact that it has the two-step process of calibration that is a composite of classes-wise correction and global normalization. The classes are corrected based on the frequency of occurrence of the respective classes in the real data and alters the predicted likelihood of the classes. Global normalization, causes the sum of probabilities of all the classes of the input to add to one. By doing this, it will reduce the overconfidence and underconfidence of predictions of the model and thus it will lead to more reasonable and strong probability estimates of cybersecurity threats detection. The CGC approach offers credible findings regarding such critical processes as the priority of alerts and response to threats by providing a combination of statistical accuracy and results that are easy to interpret. The figure presented shows that the specified methodology is more successful in comparison with the probability calibration of the conventional classifiers. This improvement is highly required in cybersecurity. More precisely-calibrated probabilities are useful in reducing false alarms, eliminating false threats and also make faster and smarter decisions. The tests using EMBER2024, CICAPT-IIoT2024 and UGRansome2024 datasets showed that the proposed method yields a better calibration under different environments. These results show that CGC is an efficient and strong device to identify multiclass threats in the complex networks.

V. CONCLUSION

This paper proposed a new method to the enhanced probabilities approximations of the multiclass cybersecurity threat apprehensions. The technique is a hybrid of normalization at the class level and normalization on a global level. This helps in the more accurate and reliable prediction of probabilities. It was tested using three cybersecurity datasets, EMBER2024, CICAPT-IIoT2024 and UGRansome2024. All datasets showed better calibration results. The crucial measures (Log Loss, Brier score and Expected Calibration Error, ECE) decreased, but the classification accuracy was not lower or even increased. To show how the uncalibrated predictions were close to the right class, Simplex plotting was performed. These

visualizations are useful in making the changes easier to understand and explain. Distorted probabilities were closer to reality. Appropriate probability estimates play an important role in cybersecurity to aid in activities like prioritization of alerts, risk rating and decision making. Calibration may be bad hence causing missed threats or false alarms. Such a method helps in mitigating such risks by enhancing the accuracy of model predictions. The suggested technique performed better with regard to accuracy and reliability compared with the common calibration procedures and baseline models. The technique minimized overconfidence and under confidence in forecasted probabilities. This contributes to the increased reliability of outputs in practice. This article demonstrates that probability calibration is worthy in high-risk areas such as cybersecurity. Future studies can investigate how to scale up the method to larger datasets, to more types of threat or alternatively, to scale up the calibration in real time to deal with changing conditions.

CRediT Author Statement
The author reviewed the results and approved the final version of the manuscript.

Data Availability
The datasets used in this study are publicly available and can be accessed through the following sources: EMBER2024 (https://emberdataset.com), CICAPT-IIoT2024 (https://www.unb.ca/cic/datasets/iiot.html), and UGRansome2024 (https://www.gti.ssr.upm.es/datasets/UGRansome). These datasets contain comprehensive cybersecurity features and labeled attack classes, which were used to train and evaluate the baseline classifiers and the proposed Class-Global Calibration (CGC) framework.

Conflicts of Interests
The authors declare that they have no conflicts of interest regarding the publication of this paper.

Funding
No funding was received for conducting this research.

Competing Interests
The authors declare no competing interests.

References
[1]. Chung, M. K., " Introduction to Logistic Regression. Journal of Statistical Analysis", 12(3), 45-58. https://doi.org/10.1016/j.jsa.2020.08.13567.
[2]. Budholiya, K., & Singh, R, "An Optimized XGBoost-Based Diagnostic System for Heart Disease Prediction",. Journal of Medical Systems, 46(2), 2012. 1-10. https://doi.org/10.1007/s10916-022-01873-2.
[3]. Biau, G, "Analysis of a Random Forests Model", Journal of Machine Learning Research, 13, 1063-1095, 2012.
[4]. Cervantes, J., "A Comprehensive Survey on Support Vector Machine"< Artificial Intelligence Review, 53(2), 1-25. https://doi.org/10.1007/s10462-020-09885-1.
[5]. Louppe, G. (2014). Understanding Random Forests: From Theory to Practice. Proceedings of the 16th International Conference on Artificial Intelligence and Statistics, 431-438. https://arxiv.org/abs/1407.7502
[6]. Tong, S., & Koller, D, "Support Vector Machine Active Learning with Applications to Text Classification", Journal of Machine Learning Research, 2, 45-66. 2001 https://jmlr.org/papers/volume2/tong01a/tong01a.pdf.
[7]. T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," Machine Learning: ECML-98, pp. 137–142, 1998, doi: 10.1007/bfb0026683.
[8]. R. J. Joyce et al., "EMBER2024 - A Benchmark Dataset for Holistic Evaluation of Malware Classifiers," Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, pp. 5516–5526, Aug. 2025, doi: 10.1145/3711896.3737431.
[9]. Ghiasvand, E., & Kha, W, "A Provenance-Based APT Attack Dataset for IIoT Environments", Journal of Cybersecurity Research, 3(2), 45-60. 2024 https://doi.org/10.1016/j.jcr.2024.07.11278.
[10]. Nkongolo, M. W., Azugo, P., & Venter, H. (2024). Ransomware Detection and Classification Using Random Forest: A Case Study with the UGRansome2024 Dataset. Journal of Cybersecurity Applications, 5(1), 22-35. https://doi.org/10.1016/j.jca.2024.04.12855.
[11]. M. Maalouf, "Logistic regression in data analysis: an overview," International Journal of Data Analysis Techniques and Strategies, vol. 3, no. 3, p. 281, 2011, doi: 10.1504/ijdats.2011.041335.
[12]. Breiman, L. (2023). Random Forests: A Retrospective and Prospective View. Machine Learning, 112(5), 1–18. https://doi.org/10.1007/s10994-023-06123-4.
[13]. T. Chen and C. Guestrin, "XGBoost," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785.

Publisher's note: The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. The content is solely the responsibility of the authors and does not necessarily reflect the views of the publisher.

ISSN (Online): 3105-9082