Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Machine Learning Framework for Cyber Risk Assessment in Cloud-Hosted Critical Data Infrastructure

[1]Veeravenkata Maruthi Lakshmi Ganesh Nerella, [2]Kapil Kumar Sharma, [3]Sarat Mahavratayajula, [4]Harish Janardhan

[1]Sr. Database Administrator
Summerfield, NC 27358, USA
ORCID:https://orcid.org/0009-0008-6149-846X

[2]Independent Researcher
Fuquay Varina, NC 27526, USA
ORCID:https://orcid.org/0009-0005-0973-6481

[3]Independent Researcher
Garner, NC 27526, USA
ORCID:https://orcid.org/ 0009-0004-4795-9777

[4]Independent Researcher
Edison, NJ 08820, USA
ORCID:https://orcid.org/0009-0004-3759-2182

ARTICLE INFO

Received: 10 July 2025

Revised: 12 Aug 2025

Accepted: 20 Aug 2025

ABSTRACT

Cloud computing has rapidly expanded to offer scalability and flexibility to current organizations, but it has also posed a new era of challenging yet dynamic security threats. With cyber threats evolving to new levels of sophistication and magnitude, organizations are finding it difficult to manage cybersecurity risks. Cybersecurity risks to the cloud-based critical infrastructure are highly sophisticated and can exact profound risks on data integrity, financial stability and operational continuity. These dynamic threats tend to be hard to foresee and thus prioritized through the conventional means of risk assessments. The research work will describe a powerful machine learning system that performs well on cyber risk to include supervised classification methods, unsupervised clustering, and anomaly detection. The steps will include preprocessing a large general global cyber threat dataset, training models with Random Forest, XGBoost, and NGBoost, model performance estimation with accuracy, precision, recall, F1-score, ROC-AUC while conducting clustering and anomaly analysis to gain further insight. Findings indicate NGBoost is the most accurately predictive (98%) and has few misclassifications, clustering reveals interesting incident patterns, and anomaly detection pinpoints high-risk sectors. The framework demonstrates especially high accuracy, interpretability, and actionable insights compared to benchmark models (CNN and KNN), which makes it a scalable and reliable proactive tool of cyber risk management. The contribution of the work is combining sophisticated feature engineering, model tuning, and multi-level analysis effectively to increase the practical cybersecurity resilience.

Keywords: Cloud, Cybersecurity, Machine Learning, Cyber Risk Scoring, Global Cybersecurity Threats dataset, XGBoost, NGBoost.

Introduction

A shared pool of computer resources may be accessed conveniently and whenever needed due to the resources management concept known as cloud computing. Cloud computing is becoming more and more commonplace while also developing quickly [1]. Cloud computing is becoming a vital pillar for businesses in contemporary digital contexts, offering a platform for scaling, adaptable, and cost-effective

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

IT resources [2]. Cloud-hosted environments are being used by enterprises to support critical operations in a variety of industries, including healthcare, finance, energy, and construction, that require continuous access to data and services [3][4]. Organisations outsource storage, networking, and computing resources to cloud service providers and gain lower capital expenditure, higher operational efficiency, and greater agility. Nevertheless, the fast emergence of cloud technologies has also accentuated the problem of cybersecurity [5]. Traditional perimeter-based security approaches are insufficient in dynamic, multi-tenant cloud environments [6], leaving infrastructures vulnerable to a wide spectrum of cyber threats [7], including data breaches, unauthorized access, insider threats, and advanced persistent threats (APTs) [8][9]. These vulnerabilities are intensified by the shared responsibility paradigm, the use of complex third-party integrations, and frequent human mistakes, including misconfigurations, which can make proactive risk evaluation and threat forecasting the keys to safeguarding sensitive data and service continuity [10][11].

Although much research has been done on cloud security, the current solutions tend to be based on static rules or risk assessment by human hands, not adapting to dynamic and advanced attack patterns [12][13]. Moreover, the existing frameworks do not include the integration of supervised, unsupervised, and anomaly detection methods regarding comprehensive cyber risk evaluation [14][15]. This creates a research gap in developing automated, intelligent, and scalable machine learning (ML) frameworks capable of assessing risk, prioritizing incidents, and detecting anomalies in real time [16]. Motivated by these challenges, this study proposes a ML-driven cyber risk assessment framework for cloud-hosted critical data infrastructures. The framework uses sophisticated methods such as supervised and unsupervised to measure the level of risk, anticipate high-risk events, profile attacks, and detect anomalies. The objectives are to enhance decision-making, improve compliance with data protection regulations, and strengthen the overall resilience of cloud-hosted infrastructures, particularly in sectors like construction, where digitalization and integration of smart technologies are increasing cybersecurity exposure. The following research contribution of this work are:

- This study proposes a comprehensive hybrid ML framework combining supervised learning (Random Forest, XGBoost, NGBoost), unsupervised clustering (PCA + K-Means), and anomaly detection (Isolation Forest) to provide an integrated approach for cyber risk assessment in cloud-hosted critical infrastructures.
- A composite risk score is developed that quantitatively evaluates threat intensity by integrating severity, frequency, and financial impact. This approach allows for dynamic, real-time risk classification into high- and low-risk incidents, enhancing decision-making.
- The framework incorporates feature importance analysis, identifying critical factors driving cyber risk. This improves model transparency, supports cybersecurity prioritization, and informs security policies for cloud-hosted infrastructures.
- Through PCA-based dimensionality reduction and K-Means clustering, the study uncovers hidden attack patterns, dominant threat types, and sector-specific vulnerabilities. These insights allow organizations to gain more context into the threat landscapes in order to tailor mitigation efforts.
- The use of an Isolation Forest enables identification of rare or new types of cyber incident, enabling early detection of anomalous attacks and preemptive risk mitigation.
- The study develops a ROC curve, accuracy, precision, recall, and F1score comparison of various ML models. The optimal modelling for predicting cyber threats is shown by this investigation, which also acts as a reference for cloud cybersecurity academics and practitioners.

This study is significant as it addresses the growing cybersecurity challenges in cloud-hosted critical infrastructures, where traditional security approaches often fail to predict and mitigate dynamic threats. By integrating supervised learning, unsupervised clustering, and anomaly detection within a single framework, the research provides a comprehensive, data-driven approach for assessing cyber risk, identifying hidden attack patterns, and detecting anomalous incidents. The novelty lies in the multi-factor risk scoring mechanism that combines severity, frequency, and financial impact, enabling dynamic classification of high- and low-risk events. Additionally, the study offers interpretability through feature importance analysis, sector-specific insights via clustering, and proactive threat detection using Isolation Forests, making it a pioneering contribution to automated, intelligent, and scalable cyber risk assessment in cloud environments.

A. *Structure of Paper*

This paper is structured as follows: Section II examines relevant research on cyberattack detection, while Section III presents the suggested approach along with a flowchart. The experimental data,

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

performance analysis, and discussion are presented in Section IV, and the study's main conclusions and future research prospects are presented in Section V.

Literature Review

This section presents research on machine learning techniques for cyber risk assessment. A cyberattack is a targeted attempt to compromise computer systems and networks in order to compromise data, impede operations, or limit access to data.

Farzaan et al. (2025) study presents a state-of-the-art cyber incident response system driven by AI that is tailored for cloud computing platforms, addressing this urgent challenge. In contrast to traditional approaches, their solution uses cutting-edge ML and AI techniques to offer accurate, scalable, and smooth interaction with platforms like Microsoft Azure and Google Cloud. Three well-known datasets, NSL-KDD, UNSW-NB15, and CIC-IDS-2017 were used to test the system and confirm its efficacy. For the categorization of network traffic, the RF model obtained 90%, 75%, and 99% accuracy, respectively, while for malware analysis, it reached 96% precision [17].

Genuario et al. (2024) proposed research establishes a baseline for the various methodologies by comparing and contrasting modern NIDSs that rely on machine learning. The suggested study contrasts two types of learning algorithms: those that are somewhat old in the field and those that are relatively new, including DL algorithms like LSTM, DNNs, and decision trees, as well as shallow learning algorithms like DT, RF, NB, LR, XGBoost, and SVM. The ensembles are also evaluated. Various datasets, including KDD-99, NSL-KDD, UNSW-NB15, IoT-23, and UNB-CIC IoT 2023, are used to assess the algorithms. The findings demonstrate that deep learning-based NIDS tools outperform all other models in identifying network abnormalities, whereas shallow learning-based tools perform worse [18].

Xiao (2024) article describes the architecture of an IoT-CTIS that uses a ML algorithm to identify security threads and malware. When it comes to memory and Internet interface vulnerabilities, all of the devices employ ML aided LR methodologies. Reduce the likelihood of system integrity issues like spoofing and DoS attacks by using the RFalgorithm. In terms of accuracy(90%), precision(90%), F-measure (88%), recall (90%), RMSE (15%), MSE (5%), TPR (89%), FNR (8%), FRP (89%), FNR (8%), security (93%), and MCC (92%). At the absolute least, it should do system testing every three months [19].

Abbas et al. (2023) objective of this research is to provide a workable plan for foreseeing the use of ML in an IndustrialCloud setting with respect to privacy and trust concerns. Validation matrices including recall, accuracy, precision, F1 scores, R.O.C. curves, and confusion matrices are used to assess the efficacy of the models. In comparison to the other models, the X.G.B. model outperformed them all in terms of accuracy (97.50%), recall (97.60%), precision (97.60%), and F1score (97.50%). This research highlights one possible application of ML algorithms to improve cloud computing security for numerous enterprises [20].

Nandhini et al. (2023) built on the detection of intrusions using three ML models: SVM, KNN, and RF. The research shows that when it comes to intrusion detection, SVM, KNN, and RF each have their own set of benefits and drawbacks. The proximity-based categorization of KNN shown to be both simple and adaptable. RF and its ensemble approach proved to be effective and flexible when confronted with complex data [21].

Alheeti et al. (2023) sophisticated intrusion detection system is recommended. On top of that, the proposed system aims to test how well the KNN can differentiate between real and altered data. The Multi-Step Cyber-Attack Dataset (MSCAD) is a trustworthy dataset that is used to identify the behavior of the new forms of attacks. In addition, the model was trained using 60% of the dataset, then tested using the remaining 40%. Furthermore, experimental results suggest that the suggested system-based KNN has the potential to enhance detection efficiency. The proposed method also reduces the number of false alarms while improving the accuracy of detections [22].

Recent research on ML-based cyber risk and intrusion detection has explored various approaches, including shallow learning algorithms like RF, SVM, KNN, and LR, as well as DL models like CNNs, DNNs, and LSTM. These techniques show decent results in identifying network abnormalities, malware, and IoT-related threats, typically scoring high on accuracy, precision and recall on known benchmark datasets. Ensemble and hybrid method further enhance the detection capabilities, whereas AI-based systems are scalable and have probabilistic interpretation. However, existing solutions have some limitations, including reliance on specific datasets, insufficient adaptability to multiple Cloud platforms, and little integration of clustering and abnormal detection to reveal underlying threat patterns. They also

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

do not provide detailed frameworks which involve both classification, exploratory analysis and anomaly detection on a single workflow. The proposed framework mitigates these gaps through its use of supervised learning, unsupervised clustering, and anomaly detection with the end result a robust, interpretable, and high-accuracy means of measuring cyber risks in cloud-hosted critical infrastructures.

Methodology

The methodology follows a structured workflow shows in Figure 1, including data collection, data cleaning, and preprocessing, feature scaling, and computation of a composite risk score. Then the dataset is further divided for training and testing on the model. Supervised learning is applied for risk classification, while unsupervised analysis is conducted for pattern discovery and clustering. Anomaly detection identifies rare or critical incidents, and further analysis provides risk insights across attack types, industries, and affected users, supported by visualizations of feature relationships, risk distributions, clusters, and anomalies.

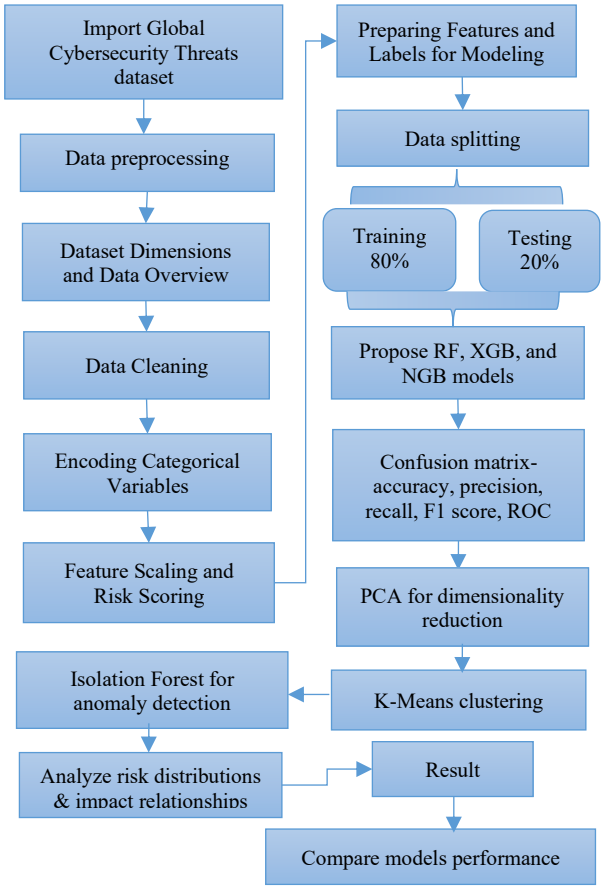

Fig. 1. Methodology Flowchart for cyber risk assessment using machine learning techniques

The following sections provide each step description that also shows in methodology and proposed flowchart:

B. *Data Collection*

The Global Cybersecurity Threats (2015–2024) dataset contains 3,000 simulated cyberattack incidents from around the world. Every entry contains information like the nation, year, attack kind (e.g., DDoS, Phishing, Ransomware), industry that was attacked, amount of money lost, amount of people impacted, attack source, exploited vulnerability, defence methods used, and time it took to resolve the issue. It is designed for cybersecurity trend analysis, threat intelligence, and machine learning applications. The dataset highlights common vulnerabilities like unpatched software and weak passwords, with industries such as IT and Banking being the most frequently targeted. It is an important tool for the comprehension of emerging global cyber threats.

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

C. *Data Cleaning and Pre-processing*

Before modeling, the dataset undergoes comprehensive cleaning. The names of the columns are standardised by making them lowercase, inserting underscores instead of spaces, and removing any unnecessary spaces for the sake of code handling and readability. Records missing critical fields like country, year, attack type, or target industry are removed to ensure data integrity. Remaining missing values in categorical fields are forward-filled, while numerical columns such as incident_resolution_time and financial_loss are imputed with their mean values. This ensures data completeness without introducing bias from arbitrary imputation. Since ML models require numerical inputs, categorical variables, attack_type, country, target_industry, and attack_source are transformed using label encoding. This process assigns each unique category a numerical value, while a dictionary of encoders is stored to enable inverse transformations later if needed. These encoded features retain the categorical meaning but are in a form that models can process efficiently.

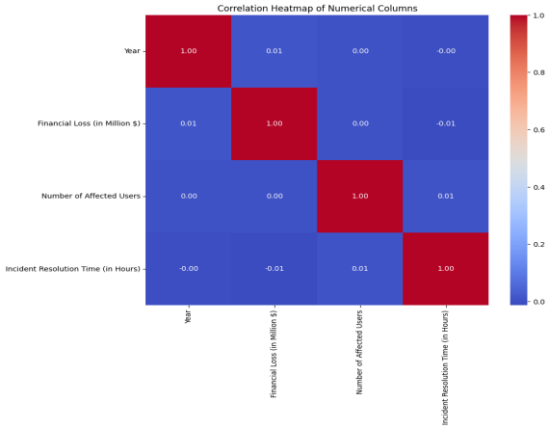

Fig. 2. Correlation Heatmap of features

Figure 2 presents the Correlation Heatmap in the relationships between key numerical variables: Year, Financial Loss (in Million $), Number of Affected Users, and Incident Resolution Time (in Hours). As shown, all features have very low correlation coefficients with each other, indicating minimal linear dependency. The values are close to zero, suggesting that no strong multicollinearity exists among the variables. Each feature may be included in modelling because of its independence; each feature adds something new to the dataset without duplicating anything else.

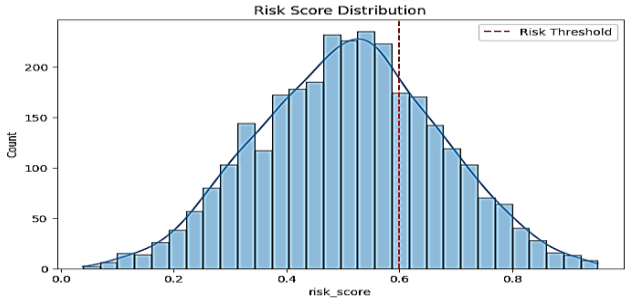

Fig. 3. Risk Score Distribution

Figure 3 displays the distribution of the computed risk_score, which integrates severity, frequency, and financial impact into a single metric scaled between 0 and 1. The histogram shows a roughly normal distribution centered around 0.5, indicating that most incidents have moderate risk levels. A red dashed line marks the threshold value of 0.6 used to classify high-risk incidents. Events falling to the right of this threshold are labeled as high risk (risk_level = 1), while those on the left are considered low risk. This split distinction will make a productive binary act with reference to cybersecurity threats.

D. *Feature Scaling, Risk Scoring, and Dataset Preparation*

The features having numerical values, incident_resolution_time, attack frequency, and financial_loss should be standardized with the Min-Max Scaler. A weighted sum is then taken to form a composite risk_score, using severity (40%), frequency (30%) and financial loss (30%) to create a basis of quantitative measure of overall threat intensity. A binary risk level label with a threshold of 0.6 is

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

applied using this score; a value of 0 indicates low risk and a value of 1 indicates high risk. A feature matrix X and a goal vector y are extracted from the dataset in order to train the model. The characteristics matrix comprises of four encoded categorical variables- attack_type_encoded, country_encoded, target_industry_encoded, and attack_source_encoded and three scaled numerical characteristics- severity_scaled, frequency_scaled and criticality_scaled. The target variable is y, which is the risk_level label, and functions as the output of predictive machine learning models. In order to test model generalization, and ensure reproducibility, the data divide is split into training (80%) and testing (20%) sets with fixed random_state=42. The model is fitting with the training set, and its performance is tested on the test set with unseen data.

E. Supervised Machine Learning Models

Three supervised classification models are trained and evaluated: Random Forest (RF)[23], XGBoost (XGB)[24], and NGBoost (NGB). The models are evaluated using accuracy, precision, recall, F1-score, ROC curves, and confusion matrices, ensuring a comprehensive assessment of both overall and class-specific predictive performance. The framework incorporates feature importance analysis, identifying critical factors driving cyber risk. This improves model transparency, supports cybersecurity prioritization, and informs security policies for cloud-hosted infrastructures. The following machine learning models are discussed below with hyperparameter tuning and evaluation:

- **Random Forest:** Hyperparameter tweaking across a predetermined parameter grid was carried out using Randomized SearchCV to optimize the RFC's performance. The search used a combination of such parameters as the number of trees (n_estimators), the depth of trees (max_depth), the number of minimum samples to split and create leaf nodes and binary use m bootstraps. The model with the highest F1score was chosen on 3-fold cross-validation on 20 randomized iterations.
- **XGBoost:** A Randomized SearchCV was used to optimize XGBoost-based classifier in order to improve its predictive power on the cybersecurity risk classification task. A wide range of parameters was set up, covered by n_estimators, learning_rate, max_depth, subsample, colsample_bytree, gamma, regalpha, and ranges sampled based on uniform or random integer distributions. The search tested 20 random combinations across 3-fold cross-validation, optimizing for the F1-score. Cybersecurity risk identification in the actual world may benefit from the provided method's high-performance, finely tailored XGBoost model.
- **NGBoost:** In this solution, NGBoost was fine-tuned manually, with an approach to scrolling through hyperparameter combinations (an n_estimators, learning_rate, minibatch_frac, natural_gradient). The F1-score was computed as a performance measure using the test data after the model was trained on the training data for each parameter set. This model's top selection was the one with the highest F1-score among competing combinations. This rigorous manual tuning provides a properly optimized NGBoost model based on the cybersecurity threat dataset.

F. Unsupervised Analysis for Pattern Discovery

Unsupervised learning practices are used to discover the latent patterns and clustering of similar cyber incidents. This starts with Principal Component Analysis (PCA) to compact the feature field down to two components to make the data simple to cluster but preserving most of the variances. Subsequently, repeated K- Means clustering are done on the PCA-transformed data over various cluster values (k 2-10). Various important parameters such as parameters like inertia (sum of squared distances) and silhouette score, and Davies-Bouldin index are measured against each k to determine the quality of clustering. To determine the best number of clusters, the Elbow Method is used on the Knee Locator algorithm that identifies the amount of clusters after which increasing cluster size has little to no additional benefit in decreasing inertia. This optimal k is visually highlighted on a plot of inertia versus cluster count, helping to select the most appropriate cluster number for the cybersecurity threat data analysis. After determining the optimal number of clusters (e.g., k=4), K-Means clustering was applied to the PCA-reduced feature space to group cybersecurity incidents based on underlying patterns. The resulting cluster labels were added to the original DataFrame for further analysis. To interpret the clustering results visually, a scatter plot of the two principal components (PC1 and PC2) was created, with points colored by cluster. This allows for intuitive understanding of how well-separated the clusters are. PCA scatter plots and t-SNE plots of the clustering results are visualized to enable the interpretation of separation. This unveils prevailing threat patterns and helps develop specific measures in cybersecurity.

G. Anomaly Detection

An Isolation Forest model is applied to detect rare or unusual attack occurrence. An Isolation Forest algorithm was used to find possible data anomalies and classify unusual cyberattack cases using feature

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

patterns. Using a contamination rate of 5%, each record was labeled as normal (1) or anomalous (-1), and the results were visualized with a count plot. This assists in identifying outlier events that could possibly be critical or rare threat scenarios. Also, it was possible to determine the top 10 high-risk industries, which were calculated based on the mean risk score per target industry. This ranking helps to visualize the sectors at the highest risk of dangerous and frequent cyber threats, prioritizing cybersecurity resources more effectively. A preview of the dataset with selected features offers a snapshot of risk levels across industries and countries over the years.

H. Risk Insights

Further investigation is conducted to plot quantities of risk scores according to attack type, target industry, attack source. Correlations between financial loss, affected users, and resolution time against risk scores present actionable intelligence on what industries and attack strategies must receive priority defensive strategies. The visualizations can give practical information about areas that are at risk, as well as put the model predictions into context.

Results and Discussion

This section shows the results of the suggested ML models to assess the cyber risks. The models were submitted in Python 3.6.5 scikit-learn library on Intel i5-8600k processor, 32GB RAM, 250GB, and 1TB HDD. RF, XGBoost, and NGBoost models were trained on the dataset, and their results were compared and measured by various measures like Acc, Prec, Rec, and F1score, at Table I. NGBoost had the most accurate overall results of 98%, Prec of 97.52%, Rec of 96.87%, and F1score of 96.67% demonstrating the overall scoring of the model that had the fewest misclassifications possible. XGBoost also performed well, and it outperformed Random Forest in all indicators, whereas Random Forest recorded the lowest values but still indicates a reasonable accuracy. These findings identify NGBoost as the most productive model in cyber risk assessment, offering a strong balance between precision and recall and offering faithful grouping in real-applications of cybersecurity.

I. Performance Matrix

There are numerous metrics that one can employ to quantify and compare the performance of ML classifiers. Accuracy assesses the extent to which samples will be correctly assigned whereas precision corresponds to the number in which the correctly predicted positive observations to the number of all predictive positives, showing reliability on detection of true risks. Recall, or sensitivity measures the power to identify all real positive results, emphasizing effectiveness at detecting abnormalities. The F1-score serves to balance between precision and recall and includes both FP and FN. The following Equations (1 to 4) of the performance measures are:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$F1-score = 2 * \frac{(precision*recall}{(precision+recall)} \quad (4)$$

The ROC curve is also used to assess the diagnostic performance of a classifier by showing the TPR against FPR at thresholds and the AUC measuring general discrimination power. High values of AUC reveal that the sensitivity and specificity are strong enough and the classification tasks are robust.

TABLE I. PARAMETERS PERFORMANCE OF PROPOSE MODELS

Measures	RF	XGB	NGB
Accuracy	95.34	97.5	98
Precision	94.55	96.11	97.52
Recall	91.14	95.57	96.87
F1-score	92.24	95.38	96.67

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

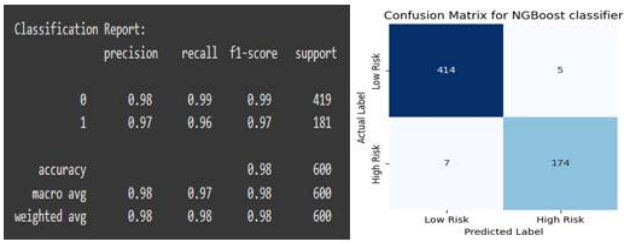

Fig. 4. Classification report and Confusion matrix of NGB classifier

Figure 4 illustrates the performance of the Naive Gradient Boosting (NGB) model in binary cybersecurity risk classification, achieving an overall accuracy of 98%. The confusion matrix shows that 414 out of 419 low-risk instances and 174 out of 181 high-risk instances were correctly classified, indicating minimal misclassifications. These results highlight the model's robustness, low error rates, and reliability for real-world cybersecurity risk prediction.

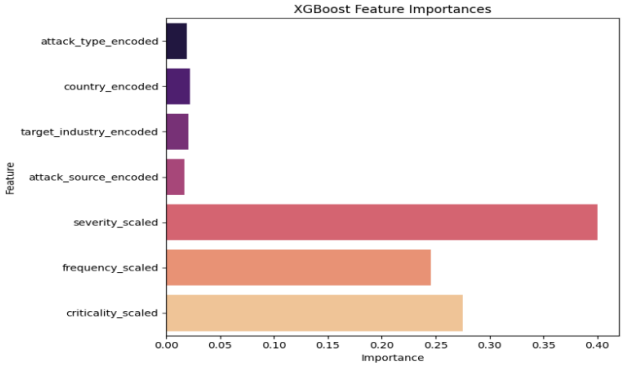

Fig. 5. Feature importance for XGB

Figure 5 illustrates the feature importance rankings from the XGBoost model, highlighting which variables most influenced its predictions. The feature severity_scaled stands out with the highest importance, followed by criticality_scaled and frequency_scaled, confirming that the severity, financial impact, and frequency of attacks are the most critical indicators of cybersecurity risk. Meanwhile, encoded categorical features like attack_type, country, target_industry, and attack_source contributed minimally to the model's decision-making. This reinforces the significance of quantifiable threat metrics over contextual attributes in effectively classifying cyber incidents.

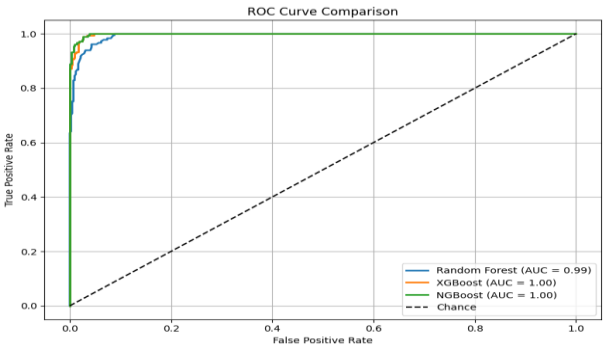

Fig. 6. ROC curve comparison of proposed models

Figure 6 compares ROC curves for RF, XGB, and NGB, showing excellent classification performance with AUC scores of 99% for RF and 100% for both XGB and NGB, indicating near-perfect risk discrimination, with XGB and NGB slightly outperforming RF.

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

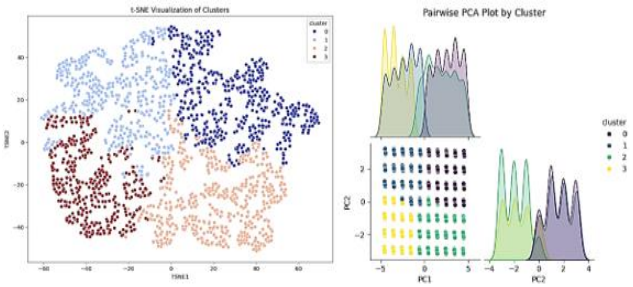

Fig. 7. Clustering results with PCA scatter and t-SNE plots

Figure 7 illustrates the clustering outcomes using PCA and t-SNE visualizations, highlighting four distinct clusters of cyber risk incidents. The t-SNE plot shows clear, well-separated groups, indicating strong non-linear structures in the data that the clustering algorithm effectively captured. In contrast, the PCA plot, based on the first two principal components, also reveals the clusters but with slightly less separation, as PCA prioritizes global variance over local relationships. The accompanying diagonal histograms in the PCA plot display feature distributions within each cluster, offering additional insights into their characteristics. Together, these visualizations validate the effectiveness of the clustering process and reveal meaningful groupings within the cyber threat landscape.

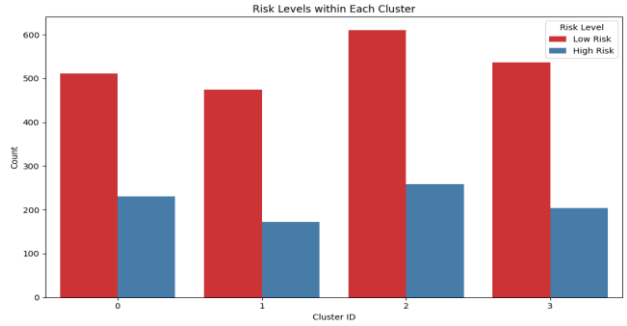

Fig. 8. Risk Levels within Each Cluster of Cybersecurity Incidents

Figure 8 shows that low-risk incidents dominate all clusters, with Cluster 2 having the highest counts overall. Clusters 0 and 2 display a higher proportion of high-risk cases, suggesting these groups may contain attack patterns or vulnerabilities linked to more severe outcomes.

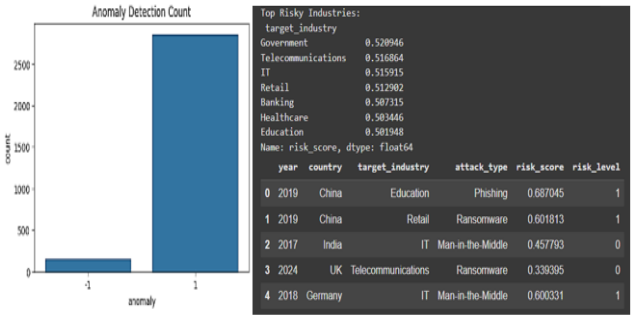

Fig. 9. Plot of Anomaly Detection Count with Summary of Top Risky Industries

Figure 9 shows most incidents as normal (1) and a small fraction as anomalies (-1). The industry summary highlights top high-risk industries by average risk score are Government (0.5209), Telecommunications (0.5166), and IT (0.5159). Example severe incidents include phishing in China's education sector (0.6870) and ransomware in China's retail sector (0.6018). The accompanying table lists example high-risk incidents, detailing the year, country, target industry, attack type, and associated risk score, underscoring the presence of severe threats like phishing, ransomware, and man-in-the-middle attacks in diverse global contexts.

Journal of Information Systems Engineering and Management
2025, 10(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

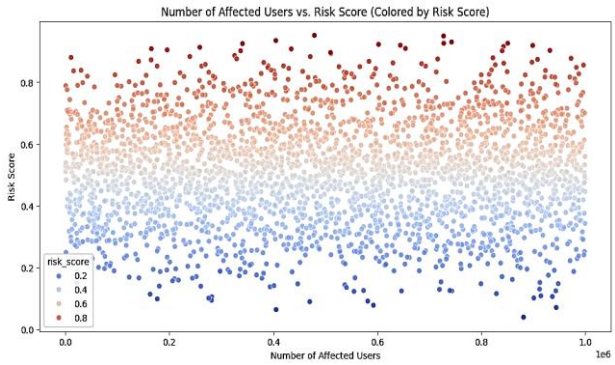

Fig. 10. Number of Affected Users vs. Risk Score (Colored by Risk Score)

Figure 10 shows the relationship among the number of affected users and the risk score, with points color-coded by risk score values. The distribution reveals that risk scores span from near 0 to above 0.9 across all scales of affected users, up to around 1,000,000. Higher risk scores (0.8–0.9, shown in red) are scattered throughout the range, indicating that severe incidents can occur regardless of the number of users impacted.

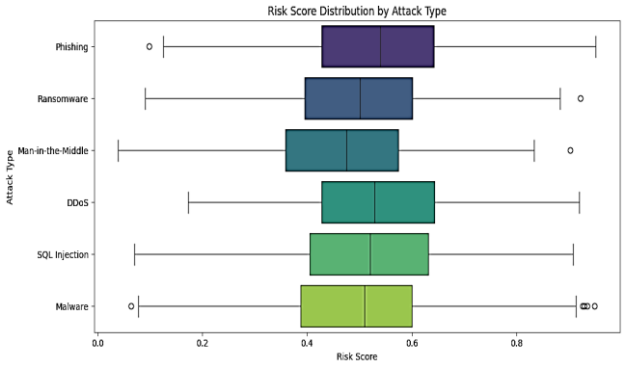

Fig. 11. Risk Score Distribution by Attack Type

Figure 11 presents the risk score distribution for different attack types using boxplots. Median risk scores are relatively similar across categories, with values clustered around 0.5. Phishing shows a narrower spread, while attacks like ransomware, DDoS, and SQL injection display wider distributions, indicating variability in severity. Outliers with risk scores near 0.9 appear in ransomware, DDoS, and malware, highlighting occasional extreme-impact incidents for these attack types.

J. Discussion

The results demonstrate that the proposed machine learning framework effectively predicts and analyzes cybersecurity risks, with all classification models achieving high performance and the NGB model leading with 98% accuracy, along with near-perfect 100% AUC scores for both NGB and XGB. Clustering analysis revealed four meaningful groups of cyber incidents, with some clusters showing higher proportions of severe cases, indicating distinct risk patterns. Anomaly detection identified a small subset of unusual, high-risk events concentrated in sectors like Government, Telecommunications, and IT, while feature-level insights showed that severe incidents could occur across all scales of affected users and varied attack types, with certain methods like ransomware and malware occasionally producing extreme impacts. Overall, the framework proves robust for accurate classification, insightful clustering, and targeted risk pattern identification in real-world cyber threat landscapes.

TABLE II. COMPARISON BETWEEN EXISTING AND PROPOSE MODELS FOR CYBER RISK ASSESSMENT BASED ON ML MODELS

Matrix	**RF**	**XGB**	**NGB**	**CNN**[25]	**KNN**[2]
Accuracy	95.34	97.5	98	97.27	84.7
Precision	94.55	96.11	97.52	-	83.1

Recall	91.14	95.57	96.87	97.78	85.1
F1-score	92.24	95.38	96.67	97.89	82.2

The comparison in Table II shows that the proposed NGB model outperforms all existing approaches for Cyber Risk Assessment, achieving the highest Acc (98%), Prec (97.52%), Rec (96.87%), and F1score (96.67%), surpassing RF, XGB, CNN, and KNN. Existing models, such as KNN, suffer from significantly lower accuracy (84.7%) and F1-score (82.2%), indicating poor generalization for complex cyber risk patterns, while CNN, though strong in recall and F1 lacks precision data and may face interpretability challenges. RF and XGB deliver competitive performance but still fall short of the NGB model, particularly in balancing precision and recall. The proposed work's advantage lies in its improved predictive accuracy, balanced classification metrics, and robustness across varied threat scenarios, addressing limitations of earlier models such as lower accuracy, incomplete performance reporting, and reduced adaptability to evolving cyber risks.

The current study's limitations primarily stem from data-related challenges, such as potential class imbalance, limited diversity in threat scenarios, and dependence on a single dataset, which may restrict the model's generalizability to real-world, dynamic cyber environments. Moreover, even though the more complex ML methods such as NGB have higher accuracy, their applicability might tiny volumes of computing memory and complicated hyperparameter optimization, hindering scalability in low-resource environments. Future efforts can concentrate on the modification of multi-source, real-time datasets, such as NSL-KDD, UNSW-NB15, CICIDS2017, TON_IoT, then apply advanced methods, such as SMOTE, ADASYN, GANs to build synthetic data, and feature selection based on mutual information or recursive feature elimination. Moreover, hybrid ML-DL systems integrating NGB, XGBoost and CNN-LSTM with explainable AI (XAI) will enhance superb interpretability, flexibility, and resistance to new and zero-day cyber threats.

Conclusion

The original machine learning architecture on cyber risk assessment of cloud hosted critical infrastructures proves to be both robust and accurate in its predictions with the Naive Gradient Boosting (NGB) model showing a use leading accuracy of 98% and nearing a perfect AUC, better than the Random Forest, XGBoost among other benchmarked models. The proposed model proves to outperform in terms of correctly identifying and evaluating cyber risks, and reducing the number of false positives, and decision-making toward security activities. The framework has used robust clustering and anomaly detection studies to provide the significant patterns of cyber incidents and the high-risk industries including Government, Telecommunications, and IT. In sum, this effort not only provides increased predictive accuracy but also provides interpretable and practical risk prioritization, proactive mitigation, and informed decision-making tools in real-world cybersecurity operations. The main contributions are the incorporation of strong features engineering, fine-tuning of the models, and benchmarking of various ML models, which demonstrates the applicability of NGB to complex cyber threat situations. Future directions should address integrating real-time threat intelligence feeds, experimenting with transformer-based deep learning models and extensions of temporal attack pre-estimation, and examining various datasets across multi- sources to augment the predication of generalizability and resilience.

References

[1] V. Prajapati, "Role of Identity and Access Management in Zero Trust Architecture for Cloud Security : Challenges and Solutions," pp. 6–18, 2025, doi: 10.48175/IJARSCT-23902.

[2] S. B. Shah, "Machine Learning for Cyber Threat Detection and Prevention in Critical Infrastructure," *J. Glob. Res. Electron. Commun.*, vol. 2, no. 2, pp. 1–7, 2025, doi: 10.5281/zenodo.14955016.

[3] D. D. Rao, S. Madasu, S. R. Gunturu, C. D'britto, and J. Lopes, "Cybersecurity Threat Detection Using Machine Learning in Cloud-Based Environments: A Comprehensive Study," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 12, no. 1, 2024.

[4] V. Shah, "Managing Security and Privacy in Cloud Frameworks: A Risk with Compliance Perspective for Enterprises," *Int. J. Curr. Eng. Technol.*, vol. 12, no. 06, pp. 1–13, 2022, doi: 10.14741/ijcet/v.12.6.16.

[5] S. Narang and A. Gogineni, "Zero-Trust Security in Intrusion Detection Networks: An AI-Powered Threat Detection in Cloud Environment," *Int. J. Sci. Res. Mod. Technol.*, vol. 4, no. 5, pp. 60–70, 2025, doi: 10.38124/ijsrmt.v4i5.542.

[6] N. Patel, "Artificial Intelligence Powered Cloud Security Detecting," 6396489, 2024

[7] S. T. Kumar, M. Shukla, N. Patel, and V. Patel, "AI Based Cyber Security Data Analytic Device," 2024.

[8] N. K. Prajapati, "Federated Learning for Privacy-Preserving Cybersecurity: A Review on Secure Threat Detection," *Int. J. Adv. Res. Sci. Commun. Technol.*, pp. 520–528, Apr. 2025, doi: 10.48175/IJARSCT-25168.

[9] D. Patel, "Leveraging Blockchain and AI Framework for Enhancing Intrusion Prevention and Detection in Cybersecurity," *TIJER – Int. Res. J.*, vol. 10, no. 6, 2023.

[10] G. Annunziata, A. Sheykina, F. Palomba, A. De Lucia, G. Catolino, and F. Ferrucci, "Security Risk Assessment on Cloud: A Systematic Mapping Study," in *Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering*, New York, NY, USA: ACM, Jun. 2024, pp. 604–613. doi: 10.1145/3661167.3661287.

[11] S. A. Pahune, P. Matapurkar, S. Mathur, and H. Sinha, "Generative Adversarial Networks for Improving Detection of Network Intrusions in IoTa Environments," *2025 4th Int. Conf. Distrib. Comput. Electr. Circuits Electron.*, pp. 1–6, 2025, doi: 10.1109/ICDCECE65353.2025.

[12] A. R. Bilipelli, "AI-Driven Intrusion Detection Systems for LargeScale Cybersecurity Networks Data Analysis: A Comparative Study," *TIJER*, vol. 11, no. 12, pp. 1–7, 2024.

[13] S. S. S. Neeli, "Critical Cybersecurity Strategies for Database Protection against Cyber Attacks," *J. Artif. Intell. Mach. Learn. Data Sci.*, vol. 1, no. 1, 2023.

[14] T. M. Chettier, V. A. K. Boyina, and S. Rangineni, "AI-Powered Risk Assessment and Compliance in Cloud Cybersecurity," *Int. J. Comput. Trends Technol.*, vol. 73, no. 3, pp. 49–56, Mar. 2025, doi: 10.14445/22312803/IJCTT-V73I3P107.

[15] R. Patel, "Automated Threat Detection and Risk Mitigation for ICS (Industrial Control Systems) Employing Deep Learning in Cybersecurity Defence," *Int. J. Curr. Eng. Technol.*, vol. 13, no. 6, 2023.

[16] V. Prajapati, "Enhancing Threat Intelligence and Cyber Defense through Big Data Analytics : A Review Study," *J. Glob. Res. Math. Arch.*, vol. 12, no. 4, pp. 1–6, 2025.

[17] M. A. M. Farzaan, M. C. Ghanem, A. El-Hajjar, and D. N. Ratnayake, "AI-Powered System for an Efficient and Effective Cyber Incidents Detection and Response in Cloud Environments," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 3, pp. 623–643, 2025, doi: 10.1109/TMLCN.2025.3564912.

[18] F. Genuario, G. Santoro, M. Giliberti, S. Bello, E. Zazzera, and D. Impedovo, "Machine Learning-Based Methodologies for Cyber-Attacks and Network Traffic Monitoring: A Review and Insights," *Information*, vol. 15, no. 11, 2024, doi: 10.3390/info15110741.

[19] P. Xiao, "Malware Cyber Threat Intelligence System for Internet of Things (IoT) Using Machine Learning," *J. Cyber Secur. Mobil.*, Dec. 2023, doi: 10.13052/jcsm2245-1439.1313.

[20] Z. Abbas and S. Myeong, "Enhancing Industrial Cyber Security, Focusing on Formulating a Practical Strategy for Making Predictions through Machine Learning Tools in Cloud Computing Environment," *Electronics*, vol. 12, no. 12, pp. 1–19, Jun. 2023, doi: 10.3390/electronics12122650.

[21] A. S, G. U. Maheswari, and S. Nandhini, "Analysis of Intrusion Detection in Cyber Attacks using Machine Learning Neural Networks," in *2023 International Conference on Sustainable Communication Networks and Application (ICSCNA)*, IEEE, Nov. 2023, pp. 1692–1696. doi: 10.1109/ICSCNA58489.2023.10370174.

[22] K. M. A. Alheeti, A. Alzahrani, O. H. Jasim, D. Al-Dosary, H. M. Ahmed, and M. S. Al-Ani, "Intelligent Detection System for Multi-Step Cyber-Attack Based on Machine Learning," in *2023 15th International Conference on Developments in eSystems Engineering (DeSE)*, IEEE, Jan. 2023, pp. 510–514. doi: 10.1109/DeSE58274.2023.10100226.

[23] N. Prajapati, "The Role of Machine Learning in Big Data Analytics: Tools, Techniques, and Applications," *ESP J. Eng. Technol. Adv.*, vol. 5, no. 2, pp. 16–22, 2025, doi:

10.56472/25832646/JETA-V5I2P103.

[24] R. Q. Majumder, "Machine Learning for Predictive Analytics: Trends and Future Directions," *Int. J. Innov. Sci. Res. Technol.*, vol. 10, no. 4, pp. 3557–3564, 2025.

[25] N. S. Kumari and N. Vurukonda, "Cyber Security Risk Assessment Framework for Cloud Customer and Service Provider," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 12, pp. 683–697, 2024, doi: 10.14569/IJACSA.2024.0151269.