ISSN: 2754-6659

Journal of Artificial Intelligence & Cloud Computing

SCIENTIFIC Research and Community

Conference Proceedings

Open Access

AI-Powered Cybersecurity Risk Scoring for Financial Institutions Using Machine Learning Techniques (Approved by ICITET 2024)

Mukund Sai Vikram Tyagadurgam[1*], Venkataswamy Naidu Gangineni[2], Sriram Pabbineedi[3], Mitra Penmetsa[4], Jayakeshav Reddy Bhumireddy[5] and Rajiv Chalasani[6]

[1]University of Illinois at Springfield, USA

[2]University of Madras, Chennai, USA

[3]University of Central Missouri, USA

[4]University of Illinois at Springfield, USA

[5]University of Houston, USA

[6]Sacred Heart University, USA

ABSTRACT

Financial institutions are confronted with an expanding range of cybersecurity risks in an increasingly digital financial environment, endangering sensitive client information and business continuity. This paper proposes an Artificial Intelligence (AI)-powered cybersecurity risk scoring model for financial institutions using machine learning (ML) techniques applied to the Lending Club dataset. The approach includes a robust preprocessing pipeline handling missing values, tokenization, normalization, one-hot encoding, and class balancing with SMOTE to enhance data quality and model fairness. Two classification algorithms, Logistic Regression (LR) and Gradient Boosting (GB), are implemented and evaluated using F1-score, recall, accuracy, and precision. The suggested models perform noticeably better than baseline models like BPSOSVMERT and Random Forest (RF). LR obtained a 99.1% F1-score, 99.6% accuracy, 99.7% precision, and 98.6% recall. GB outperformed all with 99.7% accuracy, 99.9% precision, 98.6% recall, and a 99.3% F1-score. These results highlight the effectiveness of the proposed methodology in delivering accurate and reliable cybersecurity risk predictions for enhanced decision-making in financial institutions.

***Corresponding author**

Mukund Sai Vikram Tyagadurgam, University of Illinois at Springfield, USA.

Received: April 10, 2024; **Accepted:** April 15, 2024; **Published:** April 23, 2024

Keywords: Cybersecurity Risk Scoring, Artificial Intelligence (AI), Financial Services, Smoteenn, Lending Club Dataset

Introduction

Cyber banking is the term for online banking, while cyber security is the term for procedures, methods, and technology intended to defend computer programs, networks, and data from cyberattacks. The globe is now dealing with a kind of financial terrorism known as cybersecurity threats. The hardest part of contemporary cyber banking has been protecting customers' private information [1]. One strategy to defend cyberspace from cyberattacks is cybersecurity. The goal of cybersecurity is to prevent breaches since each breach causes the target organization and its customers to suffer some kind of financial and non-financial damages [2].

Cybersecurity is a set of practices and tools designed to protect computers, networks, applications, and information from intrusions and illegal access, modification, or annihilation [3,4]. Cybersecurity risk scoring comprehensively assesses an organization's digital footprint and external threats, enabling better security posture management. This blog explores the importance of cybersecurity risk scoring and how it helps organizations mitigate cyber threats effectively [5].

The financial institutions have since been subject to stricter regulation regarding their capital sufficiency after the global financial crisis brought attention to their significance. Meanwhile, changes in the way banks function are being driven by technological advancements. At the heart of this lies AI, which has the power to completely transform financial services. It consists of a number of methods that enable computers to simulate human behavior and quickly analyze enormous amounts of data. These methods include image recognition, audio recognition, natural language processing, DL, and ML. It looks at how feasible it is to use each of these strategies in the financial services industry. In this regard, they examine risk related to credit, operations, liquidity, and reputation, all of which may negatively affect an organization's profits. AI might assist banks in reducing these risks and resolving some of

the management concerns that have been brought to light. It is determined that using AI may significantly increase the financial value of banking operations [6].

A consequence of its increasing interaction with social life, the Internet is revolutionizing how people work and learn, but it also offers more serious protection dangers. A collection of tools and procedures known as cybersecurity are intended to defend computers, networks, software, and data against intrusions and illegal access, modification, or destruction. ML/DL techniques are explained, along with a few uses each method of detecting network intrusions [7]. It focusses on ML/DL techniques and their definitions, as well as ML/DL technologies for network security [8].

Motivation and Contribution of the Paper

The motivation behind this methodology is to improve financial institutions' capacity to estimate credit risk accurately and consistently, especially by addressing the challenges posed by class imbalance in the Lending Club dataset. Traditional risk assessment methods often struggle to effectively predict the minority "Risk" class due to data imbalance, leading to biased outcomes. By integrating advanced techniques such as SMOTE for oversampling, along with robust classification algorithms like LR and GB, this approach aims to provide more accurate and equitable predictions. Financial institutions will be able to make more informed loan choices as a result, ultimately improving the efficiency and effectiveness of risk scoring systems in dynamic financial environments. The key contribution of the study is summarized below:

- **Comprehensive Preprocessing Pipeline:** The study introduces a robust data preprocessing pipeline, incorporating missing value removal, tokenization, min-max normalization, one-hot encoding, and SMOTE-based oversampling to address data quality and class imbalance issues effectively.
- **Dual-Model Approach:** The study employs both LR and GB, combining interpretability with advanced non-linear modeling to improve the predictability and precision of credit risk assessments.
- **Holistic Evaluation Metrics:** The classification models are thoroughly evaluated using a range of performance metrics, including F1-score, recall, accuracy, and precision, ensuring a comprehensive evaluation of model effectiveness.
- **Novel Application to Cybersecurity Risk Scoring:** The study extends the use of credit risk prediction algorithms driven by AI to the field of cybersecurity risk scoring in financial institutions, addressing class imbalance and providing a more reliable decision support tool for lending and cybersecurity assessments.

Justification and Novelty of the Study

Justification and novelty of the study lie in the application of a comprehensive, structured methodology for credit risk prediction using Lending Club data, which addresses both the challenges of class imbalance and the need for high predictive accuracy in financial risk assessment. The study uniquely combines advanced preprocessing techniques such as missing value removal, tokenization, min-max normalization, and one-hot encoding with SMOTE to balance the dataset, ensuring a fair representation of both "Good" and "Risk" classes. By implementing both LR and GB, the study leverages the interpretability of the former and the predictive power of the latter, providing a balanced and robust solution for credit risk prediction. This novel approach, integrating multiple steps from data preprocessing to model evaluation, creates a more reliable and effective system for credit risk scoring in financial institutions, thus advancing the state of AI-powered risk assessment in cybersecurity.

Structure of the Paper

This paper is arranged as follows: Section II of the article examines pertinent literature, Section III provides and analyses the research methods, Section IV presents and discusses the findings, and Section V concludes the study with recommendations for more research.

Literature Review

This segment offers related studies focusing on AI-powered cybersecurity risk scoring for financial institutions using ML techniques. Table I provides a concise overview of the existing literature, summarizing key approaches, methodologies, datasets used, and the outcomes achieved in various studies.

Yang et al. offer a novel technique for identifying viruses in Microsoft Word documents. To differentiate between malicious and benign MS-DOC files, an innovative approach including data extraction and conversion is developed. The analysis of MS-DOC files and the outstanding results of convolutional neural networks (CNN) in the field of feature recognition, namely image identification, serve as the driving forces behind this approach. The average accuracy rate in a simulated zero-day malware detection experiment is 94.70%, while the accuracy rate for recognition for the trial dataset is 94.09%, according to experiment results based on three CNN models [9].

Morales et al. in order to evaluate consumer data and tie it to their model data, the suggested analytical approach makes use of Bayesian networks. There are five stages to this model: 1. Analysis inputs; 2. Analysis and assessment procedures; 3. Regulations; 4. Technology architecture; 5. Output components. With an 84% prediction accuracy, this model makes it possible to determine a client's likelihood of compliance [10].

Srivastava, Agarwal and Kaur said that advancements in detection methods are necessary to identify these novel assault types. Researchers have thoroughly examined ML methods for identifying irregularities in network data. The public repositories now include new datasets. In order to identify unusual patterns in the freshly supplied dataset, the authors of this article used innovative feature reduction-based ML techniques. An 86.15% accuracy rate has been attained [11].

Sayjadah et al. present the performance evaluation of credit card default prediction. LR, part DT, and RFs are therefore used to examine the variable in credit default prediction; RF was shown to have the greatest accuracy and area under the curve. This result shows that RF best explains which factors should be included when assessing the credit risk of credit card users, with an accuracy of 82% and an area under the curve of 77% [12].

Di et al. seek to examine how well CNN performs in EEG-based person identification as the number of individuals increases. This research used a P300 speller to elicit event-related potential (ERP). The trial had thirty-three healthy participants. The results demonstrate that the CNN-based biometric system achieved a high level of accuracy (99.9%) for classifying eight classes, 99.3% for classifying ten classes, and 99.3% for classifying thirteen classes [13].

An intelligent system that allows for the detection of anomalous user behavior in online banking has been created by the research.

Because user behavior is linked to uncertainty, the system was created using fuzzy theory, which allows it to recognize user behavior and classify suspicious behavior into several intensity categories. A receiver operating characteristic curve has been used to assess the fuzzy expert system's performance, yielding a 94% accuracy rate [14].

Table 1: Summary of Literature Review of Cybersecurity Risk Scoring Using ML

Author(s)	Methodology	Dataset	Key Findings	Limitations	Future Approach
Yang et al.	CNN-based malware detection for MS-DOC files	MS-DOC files	Achieved 94.09% accuracy for malware detection; 94.70% accuracy in zero-day malware detection.	Limited to MS-DOC files; may not generalize to other file formats.	Extend detection to other file formats and improve CNN models.
Morales et al.	Bayesian networks for customer data evaluation	Customer data	84% prediction accuracy for client compliance assessment.	Focuses on customer data only; limited to a specific domain.	Explore integration with other types of data for broader applications.
Srivastava, Agarwal, Kaur	ML based on feature reduction for network traffic anomaly detection	Recently provided dataset for network traffic	Achieved 86.15% accuracy in detecting anomalies in network traffic.	May not perform well in real-time applications due to latency.	Incorporate real-time data analysis and further feature engineering.
Sayjadah et al.	Using RF , decision trees, and LR to forecast credit card default	Credit card default dataset	RF showed 82% accuracy and 77% AUC in predicting credit risk.	Focused only on credit card default prediction; limited to financial data.	Integrate additional models and features to enhance prediction accuracy.
Di et al.	CNN-based EEG-based person identification	EEG data with 33 healthy subjects	achieved classification accuracy of 99.9% for 8-class, 99.3% for 10-class, and 99.3% for 13-class.	Limited to healthy subjects; does not account for other conditions.	Expand to include diverse subject groups and conditions.
Alimolaei	A fuzzy expert system for identifying unusual activity in online banking.	Online banking user behavior data	Achieved 94% accuracy in detecting abnormal behavior using a fuzzy expert system.	Limited to online banking; may not generalize to other domains.	Extend system to detect abnormal behavior in various online platforms.

Methodology

The proposed methodology employs a structured approach for credit risk prediction using Lending Club data. Initially, the collected dataset undergoes comprehensive preprocessing, including missing value removal, tokenization, One-hot encoding and min-max normalization. SMOTE is used to solve problems with class imbalance. The prepared data is then separated into training and testing sets so that the model's performance can be objectively evaluated. LR and GB are the two classification algorithms that are used. Several performance indicators, including accuracy, precision, recall, and F1-score, are used in model evaluation to provide a comprehensive picture of classification effectiveness. This methodology ensures robust prediction capabilities while addressing the inherent class imbalance in financial risk assessment, ultimately generating reliable credit risk predictions for lending decision support. Figure 1 presents the flow diagram outlining the proposed methodology for Cybersecurity Risk Scoring for Financial Institutions.

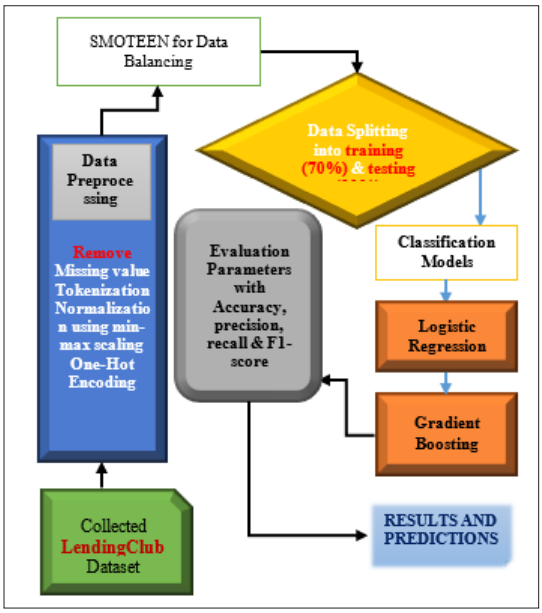

Figure 1: Flowchart for Cybersecurity Risk Scoring for Financial Institutions

The Overall Steps of the Flowchart for Financial Institutions are Provided Below:

Data Collection

The study's dataset, which includes 2,925,493 financial information from the Lending Club dataset, spanning from 2007 to 2020Q3. Various loan statuses are attributed to these records; "Charged Off," "In Grace Period," "Late (16–30 days)," "Late (31–120 days)," and "Default" denote "Risk" customers, while "Fully Paid" denotes "Good" users. "Current" status refers to ongoing payments that aren't explicitly designated as "Good" or "Risk," whilst "Issued" status refers to authorized loans that haven't yet been implemented. Users who were classified as "Does not meet the credit policy" were prohibited from taking part in the study. There are 391,882 samples classified as "Risk" and 1,497,783 samples classified as "Good." for a total of 1,889,665 samples in the collection. With an uneven ratio of 3.82, the majority class is more common in the sample. The loan status distribution is shown below in Table 2:

Table 2: Dataset from Lending Club Company from 2007 to 2020Q3 and Loan Status Distribution

Loan Status	Count	Label
Fully Paid	1,497,783	Good
Charged Off	362,548	Risk
In Grace Period	10,028	Risk
Late (16–30 days)	2,719	Risk
Late (31–120 days)	16,154	Risk
Default	433	Risk
Current	1,031,016	-
Issued	2,062	-
Does not meet the credit policy. Status: Fully Paid	1,988	-
Does not meet the credit policy. Status: Charged Off	761	-
Total	2,925,493	

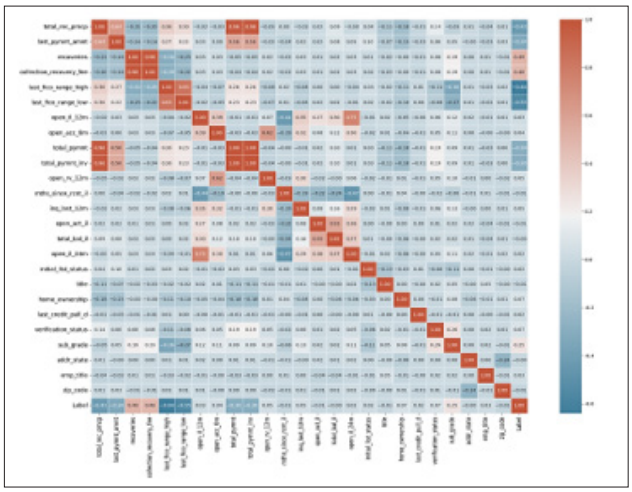

Figure 2: Correlation Matrix

Figure 2 presents the correlation heatmap showing the pairwise Pearson correlation coefficients among multiple clinical and demographic variables. Strong negative correlations are shown in deep blue, whereas strong positive correlations are shown in deep red and near-zero correlations in white or light shades. Diagonal elements have a perfect correlation of 1, while notable off-diagonal correlations include clusters among test results and age-related features. This visualization helps identify multicollinearity and relationships between variables, which is useful for feature selection in ML or statistical modeling.

Data Preprocessing

In data Lending Club dataset analysis, the process of getting raw data ready for analysis is called preprocessing by transforming it into a clean and usable format. The preprocessing steps involve Missing Value processing, Data Cleaning, Tokenization, Normalization, and One-Hot Encoding. The preprocessing steps are explained below:

Remove Missing Value: In this study, missing values were handled by removing columns with excessive missing data. The mode, or most common value, was used to impute missing values for categorical characteristics, and for continuous variables, the mean of the available data was used to fill in the missing values [15].

Tokenization: The cleansed data is converted into usable tokens via the tokenization method [16].

Normalization: Each of with a mean of 0 and an accepted variability of 1, the dataset is normalized. Data is kept within certain bounds, which reduces bias and guarantees that all members of the network get the same attention. All of the input variables in their study were normalized in the manner described below Equation (1) to be in the range [0–1] [17]:

$$x_{norm} = \frac{x_{ori} - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Where, x_{min} and x_{max} reflect the original data's lowest and highest values and x_norm and x_ori represent the original data and the normalized data, respectively [18].

One-Hot Encoding: One-hot encoding is a widely used technique for converting categorical features into numerical representations. It transforms each categorical value into a binary vector, assigning a value of 1 to the corresponding category and 0 to all other categories.

SMOTEEN for Data Balancing

Balancing the dataset is crucial for accurate classification, especially when dealing with class imbalance. This study uses SMOTEENN, a hybrid of oversampling (SMOTE) and undersampling (Edited Nearest Neighbor), to address imbalance. SMOTE generates synthetic minority samples. This combination reduces overfitting and preserves important data, outperforming other sampling methods like standalone oversampling or undersampling [19]. The synthetic samples in SMOTE are generated by Equation (2),

$$X_{new} = X_i + (X_i' - X_i) * a \quad (2)$$

Where, *Xnew* = synthetic data, X_i=instances from the minority, X_i=the k closest neighbors fromX_i, *a*= random no. from 0 and 1.

Data Splitting

In this study, data splitting is the process that helps in evaluating the model performance. They split the data into two subsets: training and testing with the following percentages, respectively: 70% and 30% [20].

Classification of Proposed Models

In this section the classification of the proposed LR and GB models are discussed below:

Classification of LR Model

Supervised learning utilizes LR as one of its popular algorithms throughout the field of ML. Supervised learning approaches use data samples to determine model parameters based on their analysis of those data samples. The model based on learning identifies actual landslide regions alongside no-landslide areas to develop forecasts for future landslide probabilities [21].

The well-known sigmoid function, the LR technique provides a chance of landslip incidence between 0 and 1. There are several sources on the specifics of the LR technique. Equation (3) provides the LR.

$$P(y = 1|x; \theta) = \frac{1}{1+e^{-\theta x^T}} \quad (3)$$

Where θ shows the parameter vector, $\theta_0,\theta_1,\theta_2,\ldots,\theta_n$, x shows feature vector, $[1,x_1,x_2,\ldots,x_n]^T$.

The parameter vector θ is repeatedly calculated with regard to a cost function using independent variables x. Equation (4) illustrates that the cost function in this investigation is binary cross entropy with L2 regularization. This cost function is used to estimate the parameters, which are changed at each iteration.

$$J(\theta) = \sum_{i=1}^{n} - \left(ylog(p) + (1-y)\log(1-p) + \frac{\lambda}{2}\sum_{j=1}^{m}\theta^2\right) \quad (4)$$

where p refers to the probability that a landslip may occur, y displays the required value, and λ is the strength of regularization.

GB Model Classification

A family of potent machine-learning methods known as GB machines has shown notable performance in a variety of real-world scenarios. They are very adaptable to the specific requirements of the application, such as learning in relation to various loss functions [22].

The base-learner models and the loss function may both be defined at will. In actuality, given a certain loss function$\psi(y,f)$ and/or a custom base-learner $h(x,\theta)$, Getting the answer to the parameter estimations might be challenging. It was suggested that a new function be chosen in order to address this $h(x,\theta_t)$ to be the most parallel to the negative gradient $\{gt\,(xi)\}N_{i=1}$ along the observed data in Equation (5):

$$g_t(x) = E_y\left[\frac{\partial\psi(y,f(x))}{\partial f(x)}|x\right]_{f(x)=\hat{f}^{t-1}(x)} \quad (5)$$

Instead of looking across the function space to find the boost increment's general solution, one may choose the new function increment with the strongest association with-$g_t\,(x)$. This makes the conventional least-squares minimization method possible job to be used in lieu of a potentially very difficult optimization work in Equation (6):

$$(\rho_t, \theta_t) = \arg\min_{\rho,\theta} \sum_{i=1}^{N}[-g_t(x_i) + \rho h(x_i,\theta)]^2 \quad (6)$$

the context of conclusion, it is able to define the GB method in its full form. The exact shape of the resulting algorithm with all of its related equations will be greatly influenced by the design choices of $\psi(y,f)$ and $h(x,\theta)$.

Evaluation Parameters

A confusion matrix is used to assess how well categorization models perform in AI-powered cybersecurity risk grading on the dataset of Lending Club. A confusion matrix was used to evaluate the categorization outcomes by comparing the predicted and actual labels. This tool compares expected results with actual labels to provide a thorough analysis of the model's prediction ability. True Positives, TN, FP, and FN are its four fundamental components, which together provide crucial information on the precision and dependability of the model's classifications. The following parameters make up the confusion matrix in the context of IDS [23]:

- **True Negative (TN):** Indicates how many normal flows were accurately identified as normal.
- **True Positives (TP):** shows the quantity of atypical flows that were appropriately identified as such.
- **False Positives (FP):** Show the quantity of typical flows that were mistakenly categorized as abnormal.
- **False Negatives (FN):** shows how many anomalous flows were mistakenly categorized as typical.

It computed the results of the proposed models in terms of precision, recall, f1-score, and overall accuracy to gauge their efficacy; these are detailed below [24]:

- **Accuracy:** indicates how accurate the model's predictions are overall. Formulated in Equation (7):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

- **Precision:** evaluates how accurate the optimistic forecasts were. The formula for mathematics shown in Equation (8):

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

- **Recall:** focusses on how well the model can detect every real positive case. expressed mathematically in Equation (9):

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

- **F1-Score:** focuses on how well the model can detect every real positive case. expressed mathematically in Equation (10).

$$F1 - Score = 2 \times \frac{Precision*Recall}{Precision+Recall} \quad (10)$$

The above performance matrices show the formulated equations of the respective parameter.

Citation: Mukund Sai Vikram Tyagadurgam, Venkataswamy Naidu Gangineni, Sriram Pabbineedi, Mitra Penmetsa, Jayakeshav Reddy Bhumireddy, et al. (2024) AI-Powered Cybersecurity Risk Scoring for Financial Institutions Using Machine Learning Techniques. Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-482. DOI: doi.org/10.47363/JAICC/2024(3)452

Result Analysis and Discussion

The experimental analysis was conducted on a high-performance computing system to ensure efficient handling of the Lending Club dataset and accurate evaluation of model performance. The system was configured with an Operating on Windows 11 Pro, it has an Intel Core i9-13900K CPU (3.0 GHz), 64 GB of DDR5 RAM, and an NVIDIA RTX 4090 GPU with 24 GB of VRAM. The evaluation results, presented in Figure 3, emphasize the strength and efficacy of the suggested ML approach for cybersecurity risk scoring in financial institutions, supporting its potential deployment in real-world financial cybersecurity applications.

Figure 3 presents a performance analysis of the proposed Accuracy, Recall, Precision, and F1-Score are the four main assessment measures for the LR and GB models. The GB model (blue bars) consistently outperforms the LR model (green bars), achieving the highest precision (99.9%) and F1-score (99.3%), while also showing a slightly higher accuracy (99.7%) compared to LR (99.6%). Both models perform equally in terms of recall (98.6%). These results demonstrate the enhanced efficacy of the suggested GB model over the LR model in detecting Instagram scam profiles.

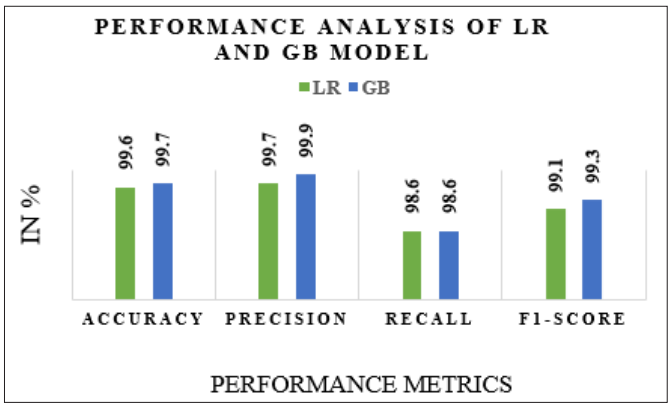

Figure 3: Performance Analysis of LR and GB Model

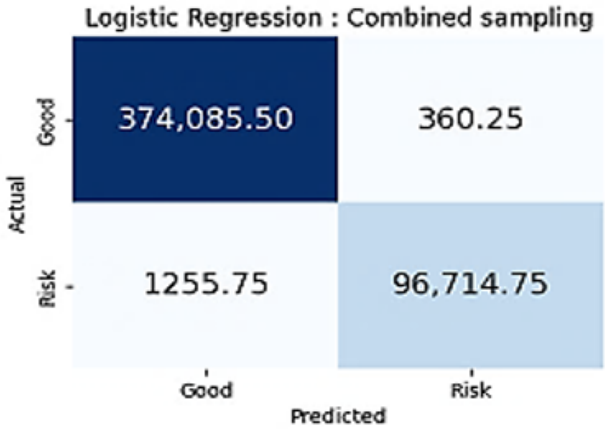

Figure 4: Confusion matrix of Logistic Regression model using combined sampling

The performance of the suggested are shown in Figure 4 LR model with combined sampling for the binary categorization of "Good" and "Risk" profiles is shown in Figure 5's confusion matrix. 374,085.50 cases were accurately identified as "Good" by the model, whereas 96,714.75 instances were classed as "Risk." 360.25 "Good" profiles were incorrectly classed as "Risk," while 1,255.75 "Risk" profiles were incorrectly classified as "Good." The efficiency of the LR model under the combined sampling technique is shown by these findings, which show good classification capabilities with extremely low FP and FN.

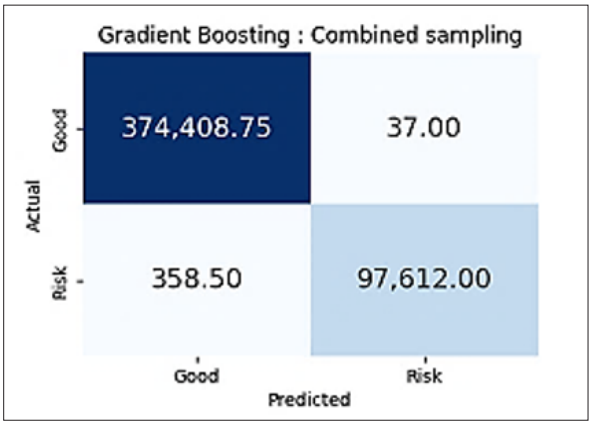

Figure 5: Confusion Matrix of Gradient Boosting Model using Combined Sampling

Figure 5 Classification results of the suggested GB model utilizing mixed sampling for prediction are shown in the confusion matrix."Good" and "Risk" profiles. The model correctly classified 374,408.75 instances as "Good" and 97,612.00 as "Risk," while misclassifying only 37.00 "Good" instances as "Risk" (false positives) and 358.50 "Risk" instances as "Good" (false negatives). These results highlight the model's high accuracy and robust performance, with very low error rates in both classes, demonstrating its suitability for imbalanced classification tasks.

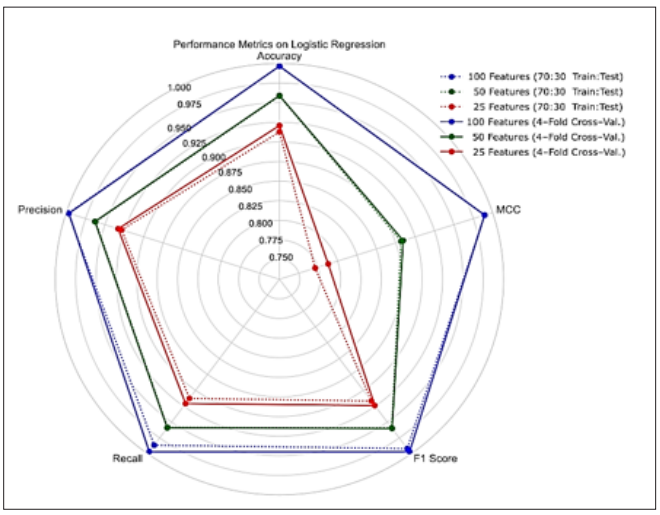

Figure 6: LR Model Performance on Five Metrics

Figure 6 radar chart illustrates the effectiveness of the LR model using various feature subsets (100, 50, and 25 features) and assessment techniques (four-fold cross-validation and 70:30 train-test split). Accuracy, precision, recall, F1 score, and Matthews Correlation Coefficient (MCC) are the measures taken into account. The model shows better overall performance with more features, particularly with 100 features, and performs slightly better with cross-validation. As the number of features decreases, performance across all metrics notably declines, especially with 25 features.

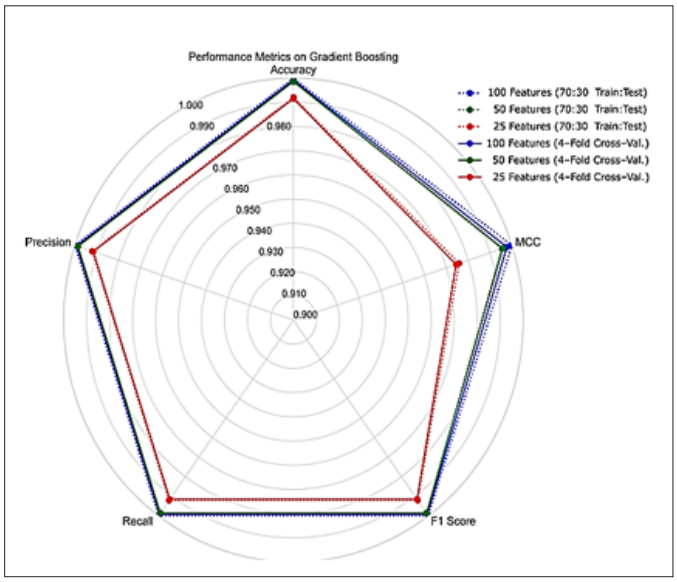

Figure 7: Gradient Boosting Model Performance on Five Metrics

The radar chart shows the performance metrics of the GB model under varying feature subsets and evaluation approaches shown in Figure 7, similar to the LR chart. GB consistently achieves high scores across all metrics, regardless of the number of features or validation method. The performance remains nearly uniform and robust, even with only 25 features, demonstrating the model's strong generalization and resilience to dimensionality reduction.

Comparative Analysis

This section represents a comparative analysis of ML algorithms integrated with combined sampling methods with relation to cybersecurity risk rating powered by AI. The evaluation includes a range of ML and DL models namely, RF, BPSOSVMERT, LR and GB, with LR and GB serving as the proposed models, while RF and BPSOSVMERT are considered baseline models. Four important performance metrics, accuracy, precision, recall, and F1-score, are used to evaluate these models [25,26]. The comparison findings, which are compiled in Table III, show the advantages and disadvantages of each strategy for precisely forecasting cybersecurity threats.

Table 3: Comparative Analysis of the Existing Models with the Proposed Models

Performance Metrics	RF [25]	BPSOSVMERT [26]	LR	GB
Accuracy	69	64	99.6	99.7
Precision	71.7	62	99.7	99.9
Recall	58.2	61	98.6	98.6
F1-Score	42	61	99.1	99.3

The comparative analysis of performance metrics demonstrates the superiority of the proposed models, shown in Table III, LR and GB, over the baseline models, RF and BPSOSVMERT. While RF and BPSOSVMERT achieved moderate performance with accuracy values of 69% and 64%, respectively, the proposed models significantly outperformed them, with LR achieving 99.6% and GB reaching 99.7%. In terms of precision, LR and GB recorded exceptionally high scores of 99.7% and 99.9%, respectively, compared to 71.7% for RF and 62% for BPSOSVMERT. The recall values for LR and GB were both 98.6%, far exceeding those of the baselines, which were 58.2% RF and 61% (BPSOSVMERT). The F1-scores further highlight the effectiveness of the proposed models, with LR at 99.1% and GB at 99.3%, while the baseline models lagged considerably at 42% (RF) and 61% (BPSOSVMERT). These outcomes validate the suggested LR and GB models' dependability and robustness for the given job.

The proposed models, LR and GB, offer several advantages, primarily due to their high accuracy, recall, precision, and F1-score. These models perform very well, which makes them ideal for situations requiring accurate and trustworthy predictions. The LR model's simplicity and interpretability provide transparent understanding of the decision-making process, while GB's Capacity to manage intricate, non-linear connections among features boosts its predictive power. Both models excel in minimizing false positives and false negatives, as indicated by their impressive precision and recall scores. The proposed models also adapt well to varying data distributions, ensuring robustness across different scenarios, which positions them as highly efficient and scalable solutions for practical applications in comparison to the baseline models.

Conclusion and Future Work

Financial institutions' security posture might be greatly improved by the growth of AI-driven cybersecurity in the industry. Using data from Lending Club, this research offers a strong AI-powered cybersecurity risk score methodology designed for financial organizations. By incorporating advanced preprocessing techniques and addressing class imbalance through SMOTE, the study ensures data quality and model fairness. The implementation of LR and GB models led to highly accurate predictions. The proposed models achieved an accuracy of 99.6% and 99.7%, respectively, significantly outperforming the baseline models RF and BPSOSVMERT. These findings confirm the potential of ML in enhancing cybersecurity risk assessment and decision support in financial environments. The current approach is limited to structured historical data and does not incorporate real-time threat signals. Additionally, it may face reduced generalizability when applied to datasets from other financial institutions or sectors. Future research will focus on integrating real-time data streams and temporal analysis to improve threat prediction capabilities. Moreover, the inclusion of ensemble techniques, advanced neural networks, and standardized ontologies for cross-institutional data sharing will be explored to further enhance model scalability and adaptability in dynamic cybersecurity landscapes [27-41].

References

1. Qiu M, Gai K, Thuraisingham B, Tao L, Zhao H (2018) Proactive user-centric secure data scheme using attribute-based semantic access controls for mobile clouds in financial industry. Futur Gener Comput Syst 80: 421-429.
2. Bhat TH, Khan AA (2015) Cybercrimes , security and challenges. International Journal of Advanced Research in Computer and Communication Engineering 4: 509-513.
3. Kolluri V (2016) A Pioneering Approach To Forensic Insights: Utilization AI for Cybersecurity Incident Investigations. Int J Res Anal Rev 3: 919-922.
4. Aftergood S (2017) Cybersecurity: The Cold War Online. Nature 547: 30-31.
5. Kaushik A, Kumar D (2019) A Cyber Risk Scoring Framework to Improve the Cyber Security Posture. Comput Secur 89.
6. Swankie G, Broby D (2019) Examining the Impact of Artificial Intelligence on the Evaluation of Banking Risk. Econ Artif Intell.

7. Kolluri V (2018) A Comprehensive Analysis on Explainable and Ethical Machine: Demystifying Advances in Artificial Intelligence. Int Res J 2: 7.
8. Yang Xin, Lingshuang Kong, Zhi Liu, Yuling Chen, Yanmiao Li, et al. (2018) Machine Learning and Deep Learning Methods for Cybersecurity. IEEE Access 6: 35365-35381.
9. Yang S, Chen W, Li S, Xu Q (2019) Approach using transforming structural data into image for detection of malicious MS-DOC files based on deep learning models. in 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference APSIPA ASC.
10. Morales EA, Ramos BM, Aguirre JA, Sanchez DM (2018) Credit Risk Analysis Model in Microfinance Institutions in Peru Through the use of Bayesian Networks. in 2019 Congreso Internacional de Innovacion y Tendencias en Ingenieria, CONIITI 2019 - Conference Proceedings, 2019.
11. Srivastava A, Agarwal A, Kaur G (2019) Novel Machine Learning Technique for Intrusion Detection in Recent Network-based Attacks. in 2019 4th International Conference on Information Systems and Computer Networks, ISCON 2019.
12. Sayjadah Y, Hashem IAT, Alotaibi F, Kasmiran KA (2018) Credit Card Default Prediction using Machine Learning Techniques. in Proceedings - 2018 4th International Conference on Advances in Computing, Communication and Automation, ICACCA 2018.
13. Di Y, An X, Liu S, He F, Ming D (2018) Using Convolutional Neural Networks for Identification Based on EEG Signals. in Proceedings - 2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics IHMSC 2018.
14. Alimolaei S (2015) An intelligent system for user behavior detection in Internet Banking. in 4th Iranian Joint Congress on Fuzzy and Intelligent Systems CFIS 2015.
15. Meng C, Liu B, Zhou L (2019) The Application Study of Consumer Credit risk model in Auto Financial Institution Based on Logistic Regression. Atlantis Paris https://www.atlantis-press.com/proceedings/msbda-19/125917036.
16. Farhan Ullah, Hamad Naeem, Sohail Jabbar, Shehzad Khalid, Muhammad Ahsan Latif, et al. (2019) Cyber security threats detection in internet of things using deep learning approach. IEEE Access 2019.
17. Vafakhah M (2013) Comparison of cokriging and adaptive neuro-fuzzy inference system models for suspended sediment load forecasting. Arab J Geosci.
18. Asadi H, Shahedi K, Jarihani B, Sidle RC (2019) Rainfall-runoff modelling using hydrological connectivity index and artificial neural network approach. Water (Switzerland) 11: 212.
19. Manju BR, Nair AR (2019) Classification of Cardiac Arrhythmia of 12 Lead ECG Using Combination of SMOTEENN, XGBoost and Machine Learning Algorithms. in Proceedings of the 2019 International Symposium on Embedded Computing and System Design, ISED 2019.
20. Gasmi H, Laval J, Bouras A (2019) Information extraction of cybersecurity concepts: An LSTM approach. Appl Sci https://www.mdpi.com/2076-3417/9/19/3945.
21. Sevgen E, Kocaman S, Nefeslioglu HA, Gokceoglu C (2019) A novel performance assessment approach using photogrammetric techniques for landslide susceptibility mapping with logistic regression, ann and random forest. Sensors (Switzerland) https://www.mdpi.com/1424-8220/19/18/3940.
22. Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. Front Neurorobot https://www.frontiersin.org/journals/neurorobotics/articles/10.3389/fnbot.2013.00021/full.
23. Teixeira MA, Salman T, Zolanvari M, Jain R, Meskin N, et al. (2018) SCADA system testbed for cybersecurity research using machine learning approach. Futur Internet https://arxiv.org/pdf/1904.00753.
24. Ha J, Kambe M, Pe J (2011) Data Mining: Concepts and Techniques.
25. Namvar A, Siami M, Rabhi F, Naderpour M (2018) Credit risk prediction in an imbalanced social lending environment. Int J Comput Intell Syst 2018.
26. Setiawan N, Suharjito, Diana (2019) A comparison of prediction methods for credit default on peer to peer lending using machine learning. in Procedia Computer Science 2019.
27. Chinta PCR (2023) The Art of Business Analysis in Information Management Projects: Best Practices and Insights 10.
28. Chinta PCR (2023) Leveraging Machine Learning Techniques for Predictive Analysis in Merger and Acquisition (M&A). Journal of Artificial Intelligence and Big Data 3: 10-31586.
29. Krishna Madhav J, Varun B, Niharika K, Srinivasa Rao M, Laxmana Murthy K (2023) Optimising Sales Forecasts in ERP Systems Using Machine Learning and Predictive Analytics. J Contemp Edu Theo Artific Intel JCETAI-104.
30. Maka SR (2023) Understanding the Fundamentals of Digital Transformation in Financial Services: Drivers and Strategic Insights. SSRN 5116707.
31. Routhu, KishanKumar, Katnapally, Niharika Sakuru, Manikanth (2023) Machine Learning for Cyber Defense: A Comparative Analysis of Supervised and Unsupervised Learning Approaches. Journal for ReAttach Therapy and Developmental Diversities 6: 1790-1803.
32. Chinta, Purna Chandra Rao, Moore, Chethan Sriharsha (2023) Cloud-Based AI and Big Data Analytics for Real-Time Business Decision-Making 36: 96-123.
33. Krishna Madhav J, Varun B, Niharika K, Srinivasa Rao M, Laxmana Murthy K (2023) Optimising Sales Forecasts in ERP Systems Using Machine Learning and Predictive Analytics. J Contemp Edu Theo Artific Intel JCETAI-104.
34. Bodepudi V (2023) Understanding the Fundamentals of Digital Transformation in Financial Services: Drivers and Strategic Insights. Journal of Artificial Intelligence and Big Data 3: 10-31586.
35. Jha KM, Bodepudi V, Boppana SB, Katnapally N, Maka SR, et al. (2023) Deep Learning-Enabled Big Data Analytics for Cybersecurity Threat Detection in ERP Ecosystems.
36. Kuraku S, Kalla D, Samaah F, Smith N (2023) Cultivating proactive cybersecurity culture among IT professional to combat evolving threats. International Journal of Electrical, Electronics and Computers 8.
37. Kalla D, Smith N, Samaah F, Polimetla K (2022) Enhancing Early Diagnosis: Machine Learning Applications in Diabetes Prediction. Journal of Artificial Intelligence & Cloud Computing. 191: 2-7.
38. Kuraku DS, Kalla D (2023) Impact of phishing on users with different online browsing hours and spending habits. International Journal of Advanced Research in Computer and Communication Engineering 12.
39. Kalla D, Kuraku S (2023) Phishing website url's detection using nlp and machine learning techniques. Journal of Artificial Intelligence 5: 145.

40. Kuraku DS, Kalla D, Samaah F (2022) Navigating the link between internet user attitudes and cybersecurity awareness in the era of phishing challenges. International Advanced Research Journal in Science Engineering and Technology 9.

41. Kuraku DS, Kalla D, Smith N, Samaah F (2023) Exploring How User Behavior Shapes Cybersecurity Awareness in the Face of Phishing Attacks. International Journal of Computer Trends and Technology https://www.ijcttjournal.org/archives/ijctt-v71i11p111.