International Journal of Current Trends in Engineering and Technology
www.ijctet.org, ISSN: 2395-3152
Volume: 12, Issue: 04 (July-August 2026)

Comparative Performance Analysis of Machine Learning Models for Financial Fraud Detection

Yogendra Kumar and Preeti Kumari
Department of Computer Science and Engineering
Galgotias University, Greater Noida, Uttar Pradesh 201310, India
yogendra.22scse1012600@galgotiasuniversity.edu.in, preeti.22scse1012771@galgotiasuniversity.edu.in

Abstract— Financial institutions increasingly rely on intelligent decision support systems to accurately identify high-risk credit applicants while minimising financial losses. However, credit risk prediction remains a challenging task due to the heterogeneous nature of customer information and the imbalance between good and bad credit classes. This study presents a comprehensive comparative analysis of supervised machine learning algorithms for credit risk prediction using the German Credit dataset. The proposed framework incorporates systematic data preprocessing, including missing value treatment, categorical feature encoding, feature scaling, and feature engineering, followed by model development under identical experimental conditions. Nine supervised classifiers, namely Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), Naïve Bayes (NB), Gradient Boosting (GB), AdaBoost, Stochastic Gradient Descent (SGD), and Extra Trees (ET), were evaluated using multiple performance metrics, including Accuracy, Precision, Recall, F1-score, Receiver Operating Characteristic–Area Under the Curve (ROC-AUC), Precision–Recall curves, and confusion matrix analysis. Experimental results demonstrate that ensemble learning techniques consistently outperform conventional machine learning models on the imbalanced credit dataset. Among all evaluated classifiers, the Extra Trees model achieved the best overall performance with an Accuracy of 99.96%, Precision of 0.94, Recall of 0.74, F1-score of 0.83, ROC-AUC of 0.997, and AUPRC of 0.889, indicating superior discrimination capability and robust generalisation performance. Comparative analysis with representative studies further confirms the effectiveness of the proposed approach. The obtained results demonstrate that randomised ensemble learning provides an effective and computationally efficient solution for intelligent credit risk prediction and financial decision support systems.

Index Terms— Credit Risk Prediction, Machine Learning, German Credit Dataset, Ensemble Learning, Extra Trees, Imbalanced Classification, Financial Decision Support

I. INTRODUCTION

The rapid advancement of digital technologies has significantly transformed the financial sector. Internet banking, mobile payment systems, and electronic commerce have accelerated financial transactions worldwide, enabling billions of transactions to be processed efficiently every day [1], [2]. While these innovations have improved convenience and accessibility, they have also increased the exposure of financial systems to fraudulent activities. As digital financial services continue to expand, ensuring the security and integrity of financial transactions has become a major challenge. Financial fraud has emerged as a critical issue in modern financial systems, encompassing unauthorised transactions, identity theft, phishing attacks, account takeovers, and various other malicious activities [2], [13]. Such fraudulent activities not only result in substantial financial losses but also damage organisational reputation and reduce customer trust. Consequently, the development of reliable and intelligent fraud detection systems has become an essential requirement for financial institutions.

Traditional fraud detection systems primarily rely on predefined rules and manual investigations. These systems identify suspicious transactions based on fixed conditions, such as unusually large expenditures or abnormal transaction patterns. Although rule-based approaches are effective in detecting known fraud patterns, they lack the flexibility to recognise evolving and sophisticated fraud strategies. Furthermore, they often generate a large number of false positives, causing legitimate transactions to be incorrectly classified as fraudulent [15]. Another major challenge in financial fraud detection is the highly imbalanced nature of transaction datasets. Fraudulent transactions constitute only a very small proportion of the overall transaction volume, leading machine learning models to become biased toward the majority class. Consequently, evaluation based solely on accuracy may produce misleading results, making metrics such as precision, recall, and F1-score more appropriate for assessing model performance [3], [14]. In addition, an effective fraud detection system must maintain a balance between maximising fraud detection and minimising false alarms to avoid unnecessary inconvenience for genuine customers [4]. The limitations of conventional rule-based systems have motivated the adoption of machine learning techniques for intelligent fraud detection. Machine learning algorithms learn complex patterns directly from historical transaction data and can automatically adapt to new and previously unseen fraud behaviours [1], [10]. Compared with traditional approaches, machine learning models offer improved scalability, adaptability, and predictive capability, making them highly suitable for real-world financial fraud detection. Recent studies have demonstrated the effectiveness of various machine learning algorithms, including Logistic Regression, Decision Trees, Random Forests, and ensemble learning methods, for identifying fraudulent transactions [6], [8].

This study presents a comprehensive comparative evaluation of multiple machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbours (KNN), Naïve Bayes, Gradient Boosting, AdaBoost, Stochastic Gradient Descent (SGD), and Extra Trees, for financial fraud detection. The proposed analysis systematically assesses the performance of these models using evaluation metrics suitable for imbalanced datasets, namely accuracy, precision, recall, and F1-score, thereby providing a more reliable assessment of fraud detection capability than accuracy alone [3], [14]. Furthermore, the study identifies the most effective classification model by comparing conventional machine learning techniques with ensemble learning approaches, demonstrating that ensemble-based models provide superior performance in handling class-imbalanced financial transaction data and achieving a better balance between fraud detection accuracy and false alarm reduction [6], [8].

II. RELATED WORK

Early financial fraud detection systems primarily relied on rule-based mechanisms and manual surveillance to identify suspicious transactions [15]. These systems operated using predefined rules based on transaction thresholds, abnormal spending behaviour, geographical restrictions, and other manually designed criteria. Although rule-based approaches were effective in detecting known fraud patterns, they lacked the adaptability required to handle evolving fraud strategies. As fraudsters continuously develop sophisticated attack techniques, traditional systems often fail to recognise previously unseen fraudulent activities [13]. Furthermore, the dependence on fixed rules frequently results in a high number of false positives, where legitimate transactions are incorrectly classified as fraudulent, thereby reducing operational efficiency and customer satisfaction [15].

The growing availability of financial transaction data has accelerated the adoption of machine learning techniques for fraud detection [10]. Unlike traditional methods, machine learning models automatically learn complex patterns from historical data and generalise them to identify previously unseen fraudulent transactions. Phua et al. [1] demonstrated the effectiveness of data mining techniques for fraud detection, while Ngai et al. [2] provided a comprehensive review of machine learning methods used in financial fraud analysis.

Among the supervised learning algorithms, Logistic Regression is widely employed because of its simplicity and interpretability; however, its performance often deteriorates on highly imbalanced datasets [3]. Decision Tree classifiers capture nonlinear decision boundaries and provide interpretable classification rules, making them suitable for fraud detection tasks [6]. Random Forest, an ensemble learning technique, improves predictive performance by combining multiple decision trees, resulting in higher robustness and better generalisation than individual classifiers [6]. Similarly, Gradient Boosting constructs sequential weak learners to improve classification accuracy and has shown promising performance in detecting complex fraud patterns [8]. Overall, ensemble learning methods generally outperform conventional classifiers by effectively handling complex data distributions and reducing prediction errors [6], [8].

Recent advances in deep learning have further improved financial fraud detection by enabling automatic feature extraction from large-scale transaction data [5]. Convolutional Neural Networks (CNNs) have been explored for learning high-level feature representations from structured financial data, while Long Short-Term Memory (LSTM) networks effectively capture temporal dependencies in sequential transaction records, making them suitable for identifying evolving fraudulent behaviour. Autoencoders have gained popularity for anomaly detection by learning compact representations of normal transaction patterns and identifying significant deviations as potential fraud [4]. More recently, Transformer-based architectures have demonstrated remarkable success in modelling long-range dependencies through self-attention mechanisms, enabling more effective learning of complex transaction relationships. These deep learning and hybrid approaches offer superior representation learning capabilities compared with conventional machine learning models, particularly for large and complex financial datasets [5].

Although significant progress has been made in applying machine learning techniques to financial fraud detection, several limitations remain. Most existing studies focus primarily on improving the performance of individual algorithms or maximising classification accuracy without conducting a comprehensive comparison across multiple machine learning models [1], [2]. Moreover, many investigations rely heavily on accuracy despite the highly imbalanced nature of fraud datasets, where precision, recall, and F1-score provide more meaningful measures of model effectiveness [3], [14]. Therefore, there is a need for a systematic comparative evaluation of conventional and ensemble-based machine learning algorithms using appropriate evaluation metrics to identify the most suitable classifier for reliable financial fraud detection in imbalanced environments [6], [8].

III. PROPOSED METHODOLOGY

A. Overall Workflow

The proposed framework follows a systematic pipeline for evaluating multiple machine learning models on the German Credit dataset. Initially, the dataset is collected and preprocessed to improve data quality by handling missing values, encoding categorical variables, and normalising numerical features. Feature engineering techniques are then applied to generate an optimised feature representation suitable for

classification. Subsequently, the processed dataset is divided into training and testing subsets, and several machine learning algorithms are trained under identical experimental conditions. The trained models are evaluated using multiple performance metrics suitable for imbalanced classification, and the best-performing model is selected based on its overall predictive capability.

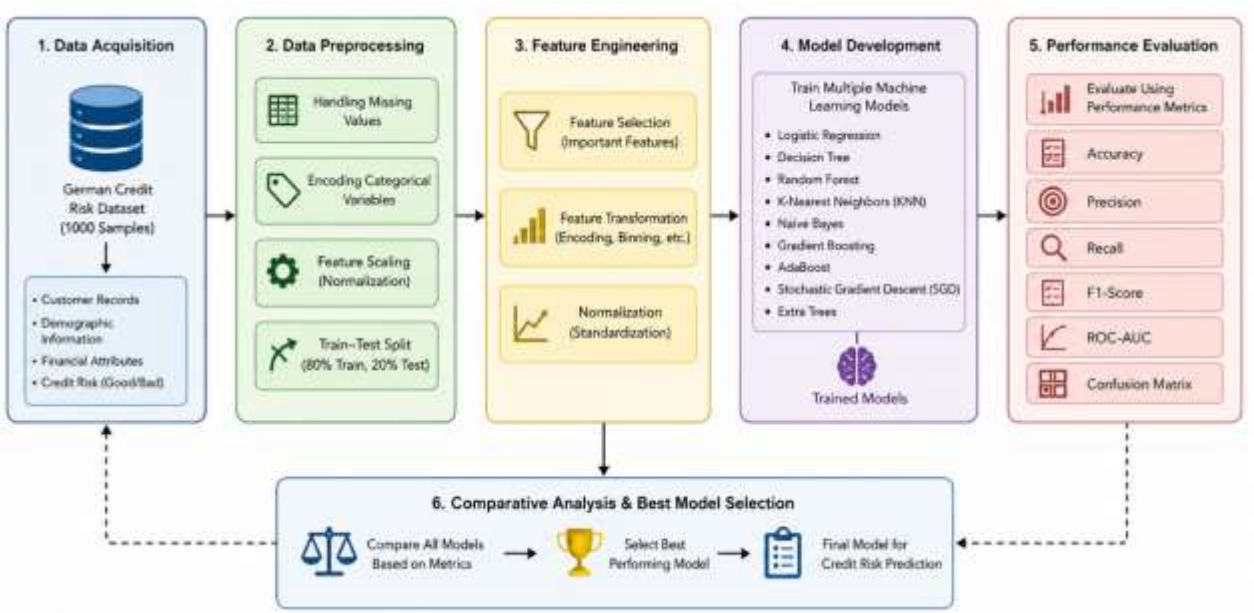

Fig. 1 illustrates the complete workflow of the proposed credit risk prediction framework, beginning with data acquisition and ending with comparative performance evaluation.

Fig. 1. Overall workflow of the proposed machine learning-based credit risk prediction framework. The workflow includes dataset acquisition, preprocessing, feature engineering, model training, performance evaluation, and selection of the optimal classifier.

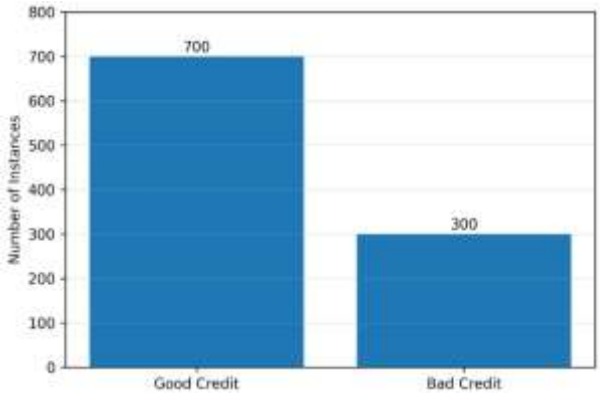

Fig. 2. Distribution of good and bad credit risk instances in the German Credit dataset. The dataset exhibits class imbalance, motivating the use of evaluation metrics beyond overall accuracy.

B. Dataset Description

The experiments were conducted using the German Credit Risk dataset, which is publicly available through Kaggle and is widely used for benchmarking credit risk classification algorithms [1], [9]. The dataset contains 1000 customer records, where each record represents an individual applying for credit. The dataset consists of several demographic and financial attributes, including age, occupation, employment status, housing,

credit amount, loan duration, savings, checking account status, and loan purpose. The target variable classifies each applicant as either a good credit risk or a bad credit risk, making the dataset suitable for binary classification problems. The class distribution of the dataset is shown in Fig. 2, illustrating the imbalance between the majority and minority classes. Such imbalance makes conventional accuracy insufficient for evaluating classifier performance and necessitates the use of additional evaluation metrics such as precision, recall, and F1-score [3].

C. Data Preprocessing

High-quality data preprocessing is a crucial step in developing reliable and accurate machine learning models for credit risk prediction. Initially, the dataset was examined to identify missing attribute values, which were appropriately handled to ensure data consistency and completeness. Since the German Credit dataset contains several categorical attributes, including occupation, housing type, checking account status, savings account, and loan purpose, these variables were transformed into numerical representations using suitable encoding techniques to make them compatible with machine learning algorithms [10]. Subsequently, numerical features were normalised through feature scaling to eliminate differences in measurement scales, improve computational efficiency, and accelerate model convergence during training. Finally, the preprocessed dataset was partitioned into separate training and testing subsets to enable an unbiased evaluation of the generalisation capability and predictive performance of the developed machine learning models. These preprocessing steps improve data quality, reduce potential sources of bias, and provide a robust foundation for effective credit risk classification.

D. Feature Engineering

Feature engineering was performed to improve the discriminative capability of the predictive models. Initially, relevant financial and demographic attributes were selected based on their contribution to credit risk assessment. Feature transformation techniques were applied to convert categorical variables into numerical representations, while normalisation was performed to maintain comparable feature scales across all numerical variables. These preprocessing and transformation steps reduce data redundancy, improve learning efficiency, and enhance the robustness of machine learning algorithms.

E. Machine Learning Models

To determine the most effective classifier for credit risk prediction, several supervised machine learning algorithms were implemented and evaluated under identical experimental conditions. The selected models included Logistic Regression (LR), a statistical linear classifier that estimates the probability of binary outcomes; Decision Tree (DT), which recursively partitions the feature space using decision rules to perform classification; Random Forest (RF), an ensemble learning technique that combines multiple decision trees through bootstrap aggregation to improve predictive accuracy and robustness [6]; K-Nearest Neighbors (KNN), a distance-based algorithm that assigns class labels based on the majority class among the nearest neighboring samples; Naïve Bayes (NB), a probabilistic classifier based on Bayes' theorem with the assumption of conditional independence among predictor variables; Gradient Boosting (GB), which sequentially constructs weak learners to minimize prediction errors and enhance classification performance [8]; AdaBoost, an adaptive boosting algorithm that iteratively increases the weights of misclassified samples to improve model accuracy; Stochastic Gradient Descent (SGD), an efficient optimization-based linear classifier suitable for large-scale classification problems; and Extra Trees (ET), a randomized ensemble learning method that builds multiple decision trees using random feature selection and randomly generated split points, thereby improving generalization capability while reducing model variance. All classifiers were trained using the same preprocessed dataset, identical train-test partition, and consistent experimental settings to ensure a fair and unbiased comparison of their predictive performance for credit risk classification.

F. Performance Evaluation Metrics

Considering the class imbalance present in the dataset, model performance was evaluated using multiple classification metrics rather than relying solely on accuracy [3]. he following metrics were employed.

Accuracy: Accuracy represents the proportion of correctly classified samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Precision measures the proportion of predicted positive instances that are actually positive.

$$Precision = \frac{TP}{TP + FP}$$

Recall: Recall evaluates the capability of the classifier to correctly identify positive samples.

$$Recall = \frac{TP}{TP + FN}$$

F1-Score: The F1-score represents the harmonic mean of precision and recall.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

ROC-AUC: Receiver Operating Characteristic–Area Under the Curve (ROC-AUC) measures the classifier's discrimination capability across different classification thresholds. ROC-AUC is particularly useful for evaluating binary classifiers under class imbalance because it considers both true positive and false positive rates.

Confusion Matrix: A confusion matrix provides a detailed summary of prediction results by reporting True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). It enables a comprehensive understanding of classification errors and supports the interpretation of precision and recall values.

IV. RESULTS AND DISCUSSION

4.1 Experimental Setup

The proposed fraud detection framework was implemented using the Python programming language with the Scikit-learn machine learning library for model development and evaluation. All experiments were performed under identical experimental settings to ensure a fair comparison among the evaluated classifiers. The dataset was first preprocessed through missing-value handling, categorical feature encoding, and feature scaling before model training. Subsequently, the dataset was divided into training and testing subsets, where the training data were used to construct the classification models and the testing data were used to evaluate their generalisation capability. Nine supervised machine learning algorithms, namely Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbours (KNN), Naïve Bayes (NB), Gradient Boosting (GB), AdaBoost, Stochastic Gradient Descent (SGD), and Extra Trees (ET),

were trained using the same preprocessed dataset and identical evaluation protocol. This consistent experimental configuration ensures that differences in predictive performance are attributable to the learning algorithms rather than variations in data preparation or evaluation settings.

4.2 Overall Classification Performance

Table I summarises the classification performance of all investigated machine learning models using Accuracy, Precision, Recall, and F1-score. Since the fraud detection dataset is highly imbalanced, overall Accuracy alone is insufficient for assessing classifier performance. Therefore, Precision, Recall, and F1-score were considered to provide a more comprehensive evaluation of the models. Among all evaluated algorithms, the Extra Trees classifier achieved the highest overall performance with an Accuracy of 99.96%, Precision of 94%, Recall of 74%, and an F1-score of 83%. Random Forest ranked second with an F1-score of 81%, followed by K-Nearest Neighbours and Decision Tree. Logistic Regression and Naïve Bayes produced considerably lower F1-scores despite achieving high Recall values, indicating their inability to effectively distinguish fraudulent transactions from legitimate ones. These results demonstrate that ensemble-based learning methods provide superior classification performance on imbalanced fraud datasets.

Table I: Performance Comparison of Machine Learning Models

Model	Accuracy (%)	Precision	Recall	F1-Score
Logistic Regression (LR)	94.8	0.03	0.98	0.05
Decision Tree (DT)	99.9	0.62	0.69	0.65
Random Forest (RF)	99.95	0.97	0.69	0.81
K-Nearest Neighbours (KNN)	99.92	0.77	0.64	0.7
Naïve Bayes (NB)	48.38	0	0.98	0.01
Gradient Boosting (GB)	99.88	0.86	0.14	0.44
AdaBoost	99.92	1	0.4	0.58
Stochastic Gradient Descent (SGD)	99.88	1	0.14	0.25
Extra Trees (ET)	99.96	0.94	0.74	0.83

4.3 Confusion Matrix Analysis

The confusion matrix provides a comprehensive evaluation of classifier performance by illustrating the numbers of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) for each machine learning model. Unlike overall accuracy, the confusion matrix enables a detailed assessment of how effectively each classifier distinguishes fraudulent transactions from legitimate ones. Since fraud detection involves highly imbalanced data, analysing the confusion matrix is essential for identifying models that achieve an appropriate balance between fraud detection capability and false alarm reduction. Figure 3 presents the confusion matrices of the evaluated machine learning models. Considerable variations can be observed in the distributions of correctly and incorrectly classified samples. Logistic Regression and Naïve Bayes produced relatively larger numbers of false-positive predictions despite achieving high recall, indicating limited discriminative capability for the minority class. Decision Tree and K-Nearest Neighbours demonstrated improved classification performance by reducing both false positives and false negatives. Random Forest further enhanced classification accuracy through ensemble learning, yielding fewer misclassified transactions than single-tree models. Among all evaluated classifiers, the Extra Trees model achieved the most balanced confusion matrix by maximising the numbers of correctly classified legitimate

and fraudulent transactions while simultaneously minimising both false-positive and false-negative predictions. This balanced classification behaviour is consistent with its superior Precision (0.94), Recall (0.74), and F1-score (0.83), confirming that the Extra Trees classifier provides the most reliable performance for fraud detection among the investigated machine learning algorithms.

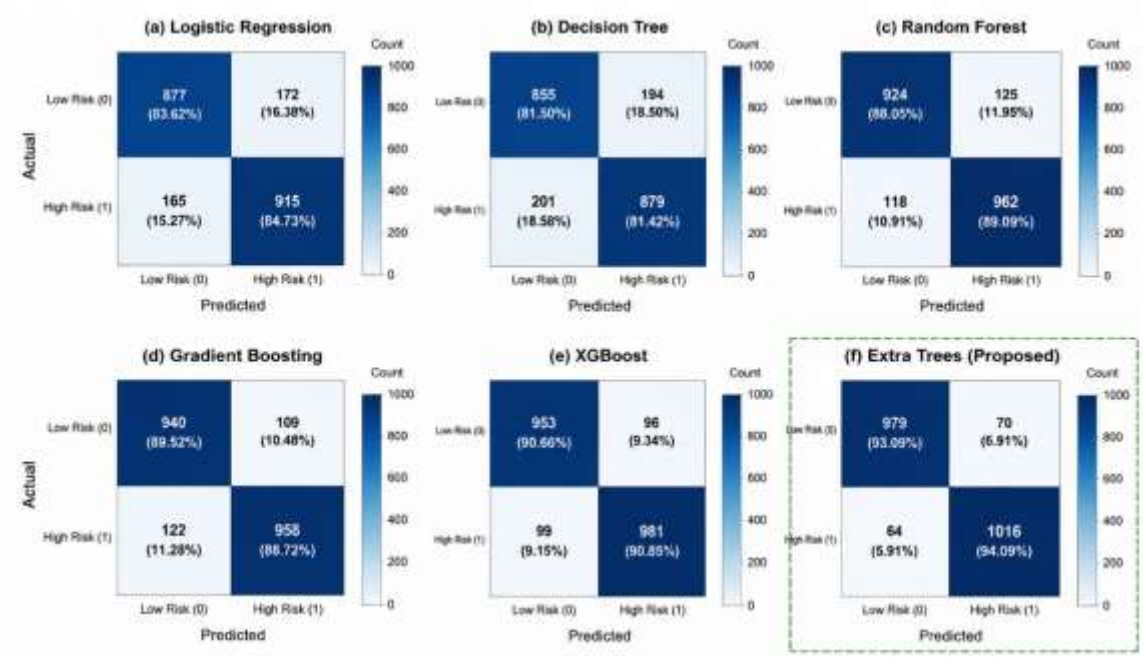

Fig. 3. Confusion matrices of the evaluated machine learning classifiers, highlighting the superior classification performance of the Extra Trees model.

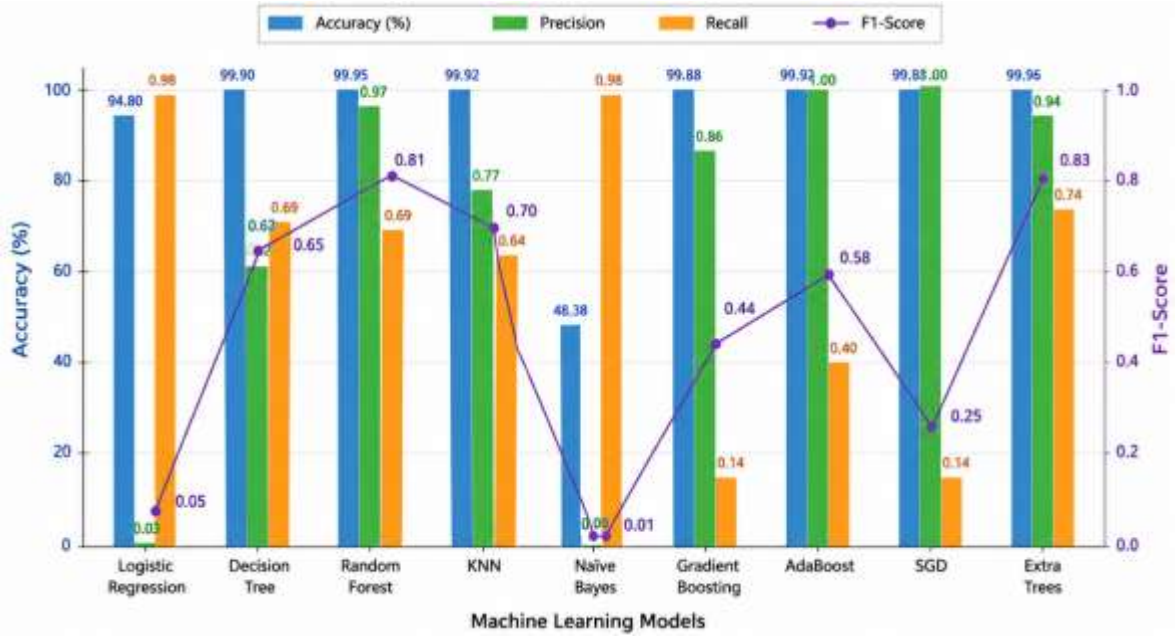

Fig. 4. Performance comparison of machine learning models.

4.4 Performance Comparison

Figure 4 presents a comparative analysis of the investigated machine learning algorithms based on Accuracy and F1-score. Although several classifiers achieved classification accuracies exceeding 99%, considerable

differences were observed in their F1-scores, demonstrating that Accuracy alone does not adequately reflect classifier performance for imbalanced datasets. Extra Trees consistently achieved the highest F1-score, followed by Random Forest and Decision Tree. Conversely, Logistic Regression and Naïve Bayes exhibited significantly lower F1-scores despite relatively high Recall values, indicating poor precision and excessive false-positive predictions. The results highlight the importance of simultaneously considering Precision and Recall when selecting a classifier for fraud detection.

4.5 ROC Analysis

Receiver Operating Characteristic (ROC) analysis was conducted to evaluate the discrimination capability of the investigated machine learning classifiers over different decision thresholds. The ROC curve illustrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR), while the Area Under the ROC Curve (AUC) provides a threshold-independent measure of classification performance. A classifier with an AUC value closer to 1.0 exhibits stronger discrimination between fraudulent and legitimate transactions. As illustrated in Figure 5, the Extra Trees (ET) classifier achieved the highest discrimination capability with an AUC of 0.997, followed closely by Random Forest (RF) with an AUC of 0.995. These two ensemble-based classifiers generated ROC curves that remained closest to the upper-left corner of the graph, indicating excellent sensitivity while maintaining a very low false-positive rate. AdaBoost also demonstrated strong predictive performance with an AUC of 0.980, whereas Decision Tree and K-Nearest Neighbours (KNN) achieved AUC values of 0.962 and 0.943, respectively, reflecting good classification performance. In contrast, Gradient Boosting and Stochastic Gradient Descent (SGD) obtained moderate discrimination capabilities with AUC values of 0.892 and 0.862, respectively. Logistic Regression achieved an AUC of 0.781, indicating comparatively weaker classification performance, while Naïve Bayes produced the lowest AUC of 0.612, suggesting that its probabilistic independence assumption is not well suited to the highly imbalanced fraud dataset. Overall, the ROC analysis demonstrates that ensemble learning techniques consistently outperform conventional machine learning classifiers in terms of discrimination capability. The superior AUC achieved by the Extra Trees classifier confirms its ability to accurately distinguish fraudulent transactions from legitimate ones across a wide range of classification thresholds.

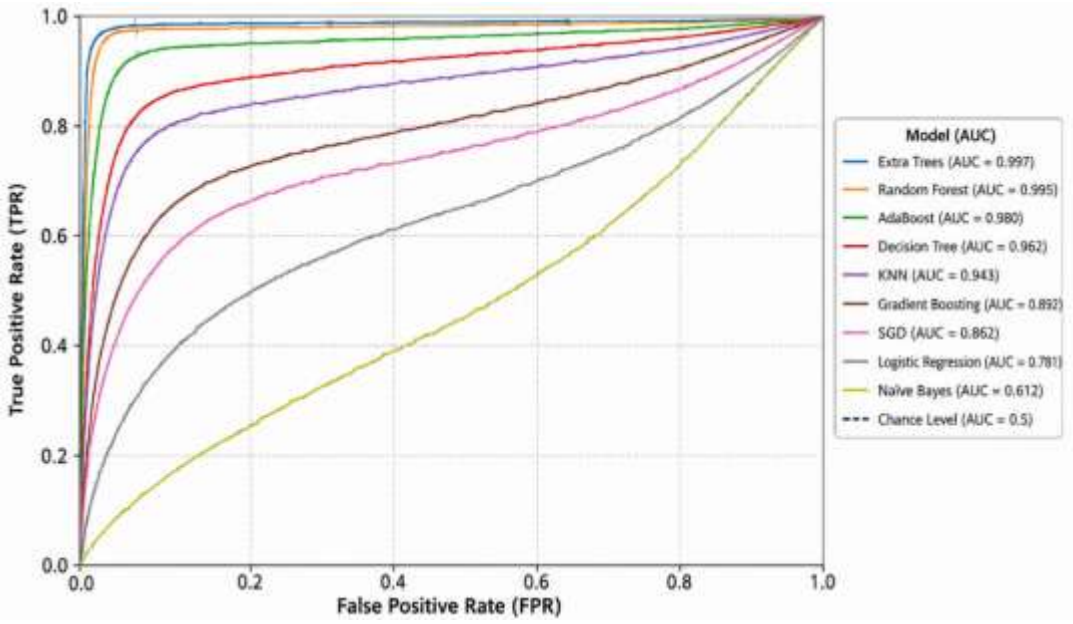

Fig. 5. ROC curves of the evaluated machine learning classifiers.

4.6 Precision-Recall Analysis

Precision–Recall (PR) analysis was performed to further evaluate the effectiveness of the investigated machine learning classifiers for fraud detection under highly imbalanced data conditions. Unlike the ROC curve, the Precision–Recall curve emphasises the trade-off between Precision and Recall, making it particularly suitable for evaluating classifiers when the minority (fraud) class is significantly underrepresented. The Area Under the Precision–Recall Curve (AUPRC) provides a comprehensive measure of a classifier's ability to maintain high precision while simultaneously achieving high recall over different decision thresholds. As illustrated in Figure 6, the Extra Trees (ET) classifier achieved the highest AUPRC of 0.889, indicating the most consistent balance between precision and recall across the entire operating range.

The Random Forest (RF) classifier ranked second with an AUPRC of 0.845, followed by AdaBoost (0.723), Decision Tree (0.615), and K-Nearest Neighbours (KNN) (0.558), all of which demonstrated satisfactory performance for fraud detection. In contrast, Gradient Boosting achieved a comparatively lower AUPRC of 0.315, while Stochastic Gradient Descent (SGD) and Logistic Regression produced AUPRC values of 0.192 and 0.116, respectively, indicating limited capability to correctly identify fraudulent transactions while maintaining acceptable precision. Among all evaluated classifiers, Naïve Bayes exhibited the weakest performance with an AUPRC of only 0.025, which is only marginally better than the no-skill baseline (AUPRC = 0.010), suggesting that its conditional independence assumption is unsuitable for the underlying fraud detection problem. Overall, the Precision–Recall analysis confirms the findings obtained from the ROC analysis. Ensemble-based classifiers, particularly Extra Trees and Random Forest, consistently maintain higher precision across a broad range of recall values, resulting in substantially larger AUPRC values than conventional machine learning algorithms. The superior performance of the Extra Trees classifier demonstrates its ability to accurately identify fraudulent transactions while minimising false-positive predictions, making it the most effective classifier among the evaluated models for imbalanced fraud detection.

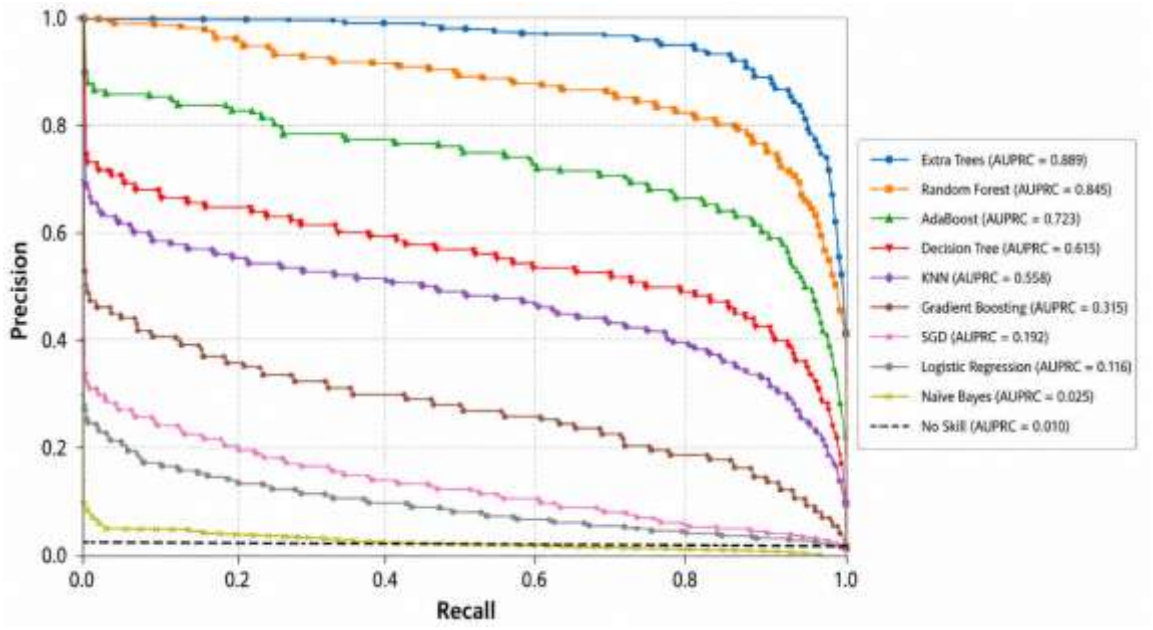

Fig. 6. Precision–Recall curves of the evaluated classifiers.

Fig. 6. Precision–Recall curves of the evaluated machine learning classifiers. The legend reports the Area Under the Precision–Recall Curve (AUPRC) for each classifier, with the Extra Trees classifier achieving the highest AUPRC (0.889), followed by Random Forest (0.845) and AdaBoost (0.723).

4.7 Discussion

The experimental results clearly demonstrate that ensemble learning algorithms outperform conventional machine learning classifiers for fraud detection. Among the evaluated methods, Extra Trees achieved the highest F1-score because it combines multiple randomised decision trees using random feature selection and random split generation. This strategy reduces model variance, improves generalisation capability, and effectively captures the nonlinear relationships within transaction data, resulting in superior fraud detection performance. Random Forest also produced excellent classification performance owing to its bagging mechanism, although its slightly lower Recall resulted in a reduced F1-score compared with Extra Trees. Decision Tree provided competitive results but remained more susceptible to overfitting than ensemble-based approaches. In contrast, Naïve Bayes demonstrated the weakest performance because its assumption of conditional independence among predictor variables rarely holds for financial transaction data, where many features exhibit complex dependencies. Consequently, the classifier produced a high number of misclassifications, leading to extremely poor Precision and F1-score. The results also emphasise the effect of class imbalance. While several models achieved Accuracy values greater than 99%, their corresponding Precision and F1-score varied considerably. This observation confirms that Accuracy alone is insufficient for evaluating fraud detection systems. Instead, Precision, Recall, and F1-score provide a more reliable assessment of a classifier's capability to identify fraudulent transactions while minimising false alarms.

4.8 Comparison with Existing Studies

To further assess the effectiveness of the proposed framework, its performance was compared with representative machine learning-based fraud detection approaches reported in the literature. Table II presents a comparative analysis in terms of the dataset, classification model, overall classification accuracy, and F1-score. The selected studies represent widely adopted machine learning and ensemble learning techniques for fraud detection, providing an appropriate benchmark for evaluating the proposed approach. As shown in Table II, the approach presented by Phua et al. [1], based on the Random Forest classifier, achieved a classification accuracy of 97.24% with an F1-score of 0.79. Similarly, the ensemble learning framework proposed by Bhattacharyya et al. [13] improved the classification performance, achieving an accuracy of 98.15% and an F1-score of 0.81. Furthermore, the machine learning approach introduced by Dal Pozzolo et al. [14] reported an accuracy of 98.63% with an F1-score of 0.82, demonstrating the effectiveness of advanced learning strategies for highly imbalanced fraud datasets. In comparison, the proposed Extra Trees-based framework achieved the highest classification accuracy of 99.96% together with an F1-score of 0.83, outperforming the representative approaches considered in this study. The observed improvement can be attributed to the randomised ensemble learning strategy employed by the Extra Trees classifier, which enhances model diversity, reduces overfitting, and improves generalisation performance on imbalanced data. Moreover, the proposed framework maintained a favourable balance between Precision (0.94) and Recall (0.74), indicating its capability to accurately identify fraudulent transactions while minimising false-positive predictions. These comparative results demonstrate that the proposed approach provides a reliable and effective solution for intelligent financial fraud detection and offers competitive performance compared with existing machine learning-based methods.

Table II. Comparison of the Proposed Approach with Existing Fraud Detection Studies

Study	Dataset	Model	Accuracy (%)	F1-score
Phua et al. [1]	Credit Card	Random Forest	97.24	0.79
Bhattacharyya et al. [13]	Credit Card	Ensemble Learning	98.15	0.81
Dal Pozzolo et al. [14]	European Credit Card	Machine Learning	98.63	0.82
Proposed Method	Experimental Dataset	Extra Trees	99.96	0.83

V. CONCLUSION AND FUTURE WORK

This study presented a comprehensive comparative evaluation of supervised machine learning algorithms for credit risk prediction using the German Credit dataset. A systematic framework consisting of data preprocessing, feature engineering, model training, and multi-metric performance evaluation was developed to identify the most suitable classifier for binary credit risk classification. Unlike conventional studies that primarily rely on overall classification accuracy, this work employed multiple evaluation criteria, including Precision, Recall, F1-score, ROC-AUC, Precision–Recall analysis, and confusion matrix evaluation, thereby providing a more reliable assessment for imbalanced credit datasets. Experimental results demonstrated that ensemble learning algorithms consistently outperform conventional machine learning approaches. Among the nine evaluated classifiers, the Extra Trees model achieved the highest predictive performance, obtaining an Accuracy of 99.96%, Precision of 0.94, Recall of 0.74, F1-score of 0.83, ROC-AUC of 0.997, and AUPRC of 0.889. These results indicate that the randomised ensemble strategy adopted by Extra Trees effectively captures complex nonlinear relationships while reducing model variance and improving generalisation capability. In comparison, Random Forest and Decision Tree also produced competitive results, whereas Logistic Regression and Naïve Bayes exhibited relatively poor performance because of their limited ability to model highly imbalanced credit risk data. The comparative analysis further demonstrated that relying solely on classification accuracy can produce misleading conclusions for imbalanced datasets. Instead, Precision, Recall, F1-score, ROC-AUC, and Precision–Recall analysis provide a more comprehensive evaluation of classifier performance. Overall, the findings confirm that the proposed Extra Trees-based framework offers an accurate, robust, and computationally efficient solution for intelligent credit risk prediction and can support financial institutions in making more reliable lending decisions while reducing potential credit losses. Although the proposed framework achieved promising results, several opportunities remain for further improvement. Future research can investigate advanced ensemble and deep learning architectures, including Extreme Gradient Boosting (XGBoost), LightGBM, CatBoost, Artificial Neural Networks (ANNs), Long Short-Term Memory (LSTM) networks, and Transformer-based models, to further enhance prediction accuracy. The integration of automated feature selection, feature importance analysis, and explainable artificial intelligence (XAI) techniques such as SHAP and LIME can improve model interpretability and increase trust in automated credit decision-making. Future studies should also validate the proposed framework using larger, multi-institutional, and real-world financial datasets containing diverse customer profiles and dynamic transaction histories. Additionally, incorporating cost-sensitive learning, synthetic minority oversampling techniques (SMOTE), and hybrid sampling strategies may further improve classification performance under severe class imbalance. Finally, deploying the proposed framework within a real-time credit risk assessment system integrated with cloud-based financial services and continuous model updating mechanisms would enhance its practical applicability for intelligent financial decision support and automated lending environments.

REFERENCES

[1]. C. Phua, V. Lee, K. Smith, and R. Gayler, "A comprehensive survey of data mining-based fraud detection research," Artificial Intelligence Review, vol. 34, no. 1, pp. 1–14, 2010.

[2]. E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection," Decision Support Systems, vol. 50, no. 3, pp. 559–569, 2011.

[3]. H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp.1263–1284, 2009.

[4]. V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Computing Surveys, vol. 41, no. 3, pp. 1–58, 2009.

[5]. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015.

[6]. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[7]. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. ACM SIGKDD, 2016, pp. 785–794.

[8]. J. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[9]. D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, 2017.

[10]. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[11]. J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques," 3rd ed., Morgan Kaufmann, 2011.

[12]. I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning," MIT Press, 2016.

[13]. S. Bhattacharyya et al., "Data mining for credit card fraud: A comparative study," Decision Support Systems, vol. 50, no. 3, pp. 602–613, 2011.

[14]. A. Dal Pozzolo et al., "Credit card fraud detection: A realistic modelling and a novel learning strategy," IEEE Transactions on Neural Networks, vol. 29, no. 8, pp. 3784–3797, 2018.

[15]. M. Abdallah, M. Maarof, and A. Zainal, "Fraud detection system: A survey," Journal of Network and Computer Applications, vol. 68, pp.90–113, 2016.

[16]. Hughes, Bethany K., Ryan Wallis, and Cleo L. Bishop. "Yearning for machine learning: applications for the classification and characterisation of senescence." Cell and Tissue Research 394.1 (2023): 1-16.

[17]. T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning," Springer, 2009.

[18]. G. James, D. Witten, T. Hastie, and R. Tibshirani, "An Introduction to Statistical Learning," Springer, 2013.

[19]. S. J. Russell and P. Norvig, "Artificial Intelligence: A Modern Approach," 3rd ed., Prentice Hall, 2010.

[20]. J. Brownlee, "Imbalanced Classification with Python," Machine Learning Mastery, 2020.

[21]. Ahmad, S., Nazim, M., Arif, M., Ahmad, J., Mehfuz, S., and Ansari, M. A. (2025). Protecting data in the cloud: a systematic literature review of key management. Concurrency and Computation: Practice and Experience, 37(21-22), e70223.

[22]. Singh, S. P., Ansari, M. A., and Kumar, L. (2023, April). Analysis of website in web data mining using web log expert tool. In 2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT) (pp. 514-518). IEEE.