InfoTech Spectrum: Iraqi Journal of Data Science (IJDS)
ISSN: 3007-5467, Vol 2(No.2), 44-60 (2025)
DOI: https://doi.org/10.51173/ijds.v2i2.22
ijds.mtu.edu.iq

LOAN REPAYMENT DEFAULT PREDICTION USING SUPERVISED MACHINE LEARNING TECHNIQUES ON FINANCIAL DATA

Muna Abdulmunem Othman [1], Ali Raad [1 *], Ahmad Ghandour [2]

[1] *Faculty of Sciences and Arts, Islamic University of Lebanon (IUL), Lebanon*
[2] *Faculty of Engineering, Islamic University of Lebanon (IUL), Lebanon*

** Corresponding author E-mail: ali.raad@iul.edu.lb (Ali Raad)*

RESEARCH ARTICLE

ARTICLE INFORMATION

SUBMISSION HISTORY:
Received: 7 February 2025
Revised: 22 April 2025
Accepted: 17 May 2025
Published: 30 June 2025

KEYWORDS:
Machine Learning;
Loan Prediction;
Weka;
Financial Risk Assessment;
Supervised Learning;

ABSTRACT

With the enhancement of technology facilitating the expansion of businesses and thoughts, more and more people are applying for loans for personal or business use. However, banks have limited assets, which limit the amount of loans that can be granted. Identifying the right persons to grant loans to can be a time-consuming process. Banks seek to grant loans to individuals who can repay the loan on time, enabling the bank to obtain maximum profits. This work aims to solve the loan default problem with minimum costs to banks. This work consists of five main stages: pre-processing, feature extraction, machine learning techniques, evaluation models, and performance analysis to select the best machine learning models. Then, two datasets with different features are used. The first dataset has five features, and the second contains eighteen features. We are splitting the datasets into various training percentages (40%, 50%, 60% and 70%). The rest of the dataset is used for testing using only the Weka application. KNN is applied with different cross-validations, such as 15, 10, and 5, and different numbers of nearest neighbours (1, 5, 10, and 15). For the first dataset, the highest accuracy is 97.47% with two cross-validation values, 15 and 10, in the 10 nearest neighbours. The KNN was also implemented on the second dataset to compute the highest accuracy, 88.21% in three cross-validation values (15, 10, and 5) with the 15 nearest neighbours. Then, logistic regression is applied to compare the results of the correct classification value computed at the highest value of 96.93% with the (70% training set for the first dataset. The highest accuracy was obtained at 88.32% after splitting the second dataset (40%) for training and the rest for testing.

1. INTRODUCTION

The rapid evolution of technology has accelerated decision-making processes, resulting in a surge in credit card activity in countries like the United States. Fair, Isaac & Co. (FICO) was founded in 1956 to assist in consumer credit evaluation, with computer-based credit application processing emerging in the 1960s [1]. By 1975, multivariate discriminant analysis gained formal approval for credit evaluation after the passage of the Equal Credit Opportunity Act [2]. Similarly, in the United Kingdom, the first credit card, Barclaycard, was introduced in 1966 [3].

Despite technological advances, accurately assessing loan applications remains a significant challenge for financial institutions. Traditional methods often struggle to predict loan defaults reliably, especially as application volumes rise and more diverse borrower profiles emerge. In addition, Rural lending, which is essential for global agricultural production, is also significantly affected by the accuracy of credit assessments. In rural areas, where financial records are often limited and income is seasonal, inaccurate credit evaluations can lead to unjust loan denials or increased default rates. This highlights the need for adaptive, data-driven risk models to ensure fair and sustainable rural finance [4]. Human biases and limited model generalizability further complicate decision-making, leading to financial losses and inefficiencies. Research proposes developing a machine learning model aimed at accurately identifying non-performing loans. The study seeks to enhance credit risk assessment by leveraging transfer-based machine learning techniques while minimizing human bias and increasing decision accuracy. The model is evaluated using key performance metrics—precision, recall, and F1-score—and compared against existing

DOI: https://doi.org/10.51173/ijds.v1i2.22

approaches to validate its effectiveness. Machine learning (ML) has emerged as a vital tool in data science, offering transformative opportunities for financial forecasting and decision-making [5]. This study specifically focuses on building a system that detects defaulting borrowers through a structured approach:

- Developing a transfer learning-based ML framework tailored for banking datasets to predict loan defaults.
- Conducting a comparative evaluation with previous studies based on standard classification metrics.

The remainder of this manuscript is structured as follows: Section 2 discusses the related work; Section 3 details the methodology; Section 4 presents the proposed system and experimental results; Section 5 compares the results with existing approaches; and Section 6 concludes the study.

2. LITERATURE REVIEW

Recent studies in credit default prediction indicate the increasing reliance on artificial intelligence and machine learning technologies as pivotal tools for improving the accuracy of credit assessments and accelerating the loan-granting process. However, with the proliferation of digital transformation and intensifying competition between financial institutions, the introduction of advanced predictive models acts as a mitigator in reducing the risks incurred from non-performing loans and reaching financial solvency for banks. Some of the most successful models that researchers have investigated are logistic regression, support vector machines, decision trees, random forests, k-nearest neighbour algorithm, multi-layer perceptron, deep learning, etc.[6] Although it is a simple method to interpret results, logistic regression performance in many research studies is low in predicting complex or non-linear data and is sometimes inaccurate [7]. Researchers have tried to transform this model to banking data to reflect banking cost-sensitivity by developing cost-sensitive algorithms on actual financial costs, i.e., by embedding a state-dependent cost matrix into logistic regression [8], which achieves remarkable result improvements and financial saving gains.

On the other hand, Support Vector Machine (SVM) has good customer classification results, especially in biclass data among defaulters and non-defaulters. And it has been demonstrated to work well with high-dimensional data and has produced accurate results when using multiple kernels (e.g., polynomial kernels ([9])). However, due to the need for manual tuning parameters and vulnerability to data distribution, the effectiveness of SVM is limited when faced with severely rare or heterogeneous data. Decision trees (DT) were listed for their simple interpretation, efficiency as exploratory instruments, and high accuracy in several studies running to segment customers by creditworthiness [10]. However, one fatal shortcoming of such a model is that it overfits and its generalizability is compromised. For this reason, boosted classifiers like AdaBoost have been used or are being embedded in the Random Forest to give better performance. Although simple and without assumption about data distribution, both being its main advantages, the K-nearest Neighbor (KNN) algorithm is proven not to work well with large datasets, more so with an increased computation overhead and lower accuracy for noisy/ non-standard features [11]. It has been used in some studies for limited performance comparisons, but it rarely yields optimal results. Random Forests (RF) have shown significant superiority in many studies. They rely on combining multiple decision trees trained on subsamples, which improves model accuracy and reduces bias. They have consistently performed superior in many experiments compared to DT or SVM [12]. However, some other studies have indicated that RF may not always be optimal, especially when dealing with large and complex datasets, where models such as XGBoost have emerged as more efficient alternatives in terms of speed and accuracy [13]

It is also worth noting that many studies have focused on the pre-model building stage, i.e., preprocessing, given its significant impact on the quality of results. It has been observed that banking data often suffers from class imbalances, missing values, and noise, requiring techniques such as cleaning, data normalization, outlier handling, and dimensionality reduction. Studies such as [14] and [15] have demonstrated the importance of selecting the most meaningful attributes of

credit behavior, which has contributed to improving the performance of subsequent models. Finally, some studies have sought to adopt hybrid models that combine the advantages of more than one algorithm, or use deep learning techniques with RBF or artificial neural network-based structures. These models have achieved promising results, especially in complex cases that are difficult to represent mathematically using traditional algorithms [16], [17]. Despite substantial progress, several critical gaps remain in the current literature:

- **Conflicting findings** regarding the most effective ML algorithm for credit risk prediction (e.g., SVM vs. RF vs. XGBoost) limit the establishment a standard, universally accepted approach.
- **Limited focus on transfer learning**: While traditional supervised models have been extensively explored, few studies have leveraged transfer-based machine learning frameworks for banking datasets.
- **Underexplored cost-sensitive evaluation metrics**: Although cost-sensitive algorithms have been proposed, their application across diverse models remains insufficiently validated, especially for highly imbalanced datasets standard in loan default prediction.
- **There is a lack of comparative studies** that holistically evaluate models using consistent metrics such as precision, recall, and F1-score across different datasets.

These limitations restrict our theoretical understanding and the practical deployment of intelligent loan evaluation systems in real-world financial environments. In response to these gaps, this study proposes a novel transfer-based machine learning approach to enhance default prediction accuracy while reducing human biases. It systematically evaluates performance using standardized metrics and benchmarks results against leading techniques from existing literature, offering a comprehensive and state-of-the-art solution for modern financial institutions.

3. RESEARCH METHODOLOGY

Fig. 1 shows the proposed model to predict accurate loan repayment and default behavior using machine learning. The approach starts with the data preprocessing phase, which is the most crucial part to manage the data quality we are using. Two records were connected from each of two databases: one table with customer identifiers (Dataset 1) and one with the details of their loans (Dataset 2). These two datasets were merged with matching IDs (ID–Loan ID).

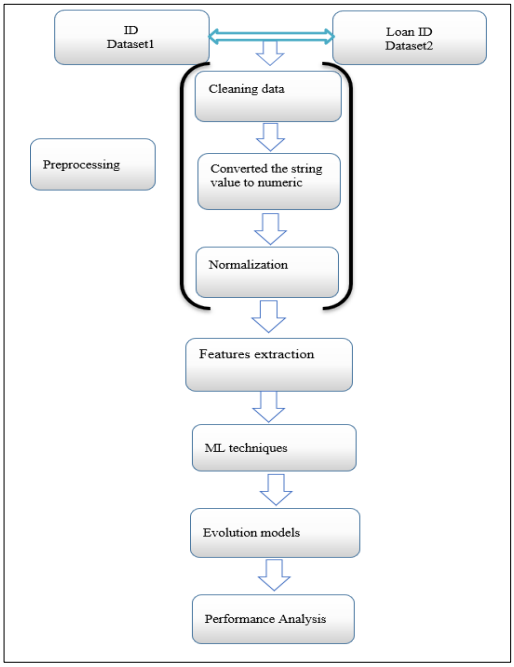

Figure 1: Proposed system block diagram

Preprocessing for this data includes a series of tasks such as cleaning the data from missing values and duplicates, transforming to a similar data format, and finally converting text to numbers

using techniques like One-Hot Encoding and Label Encoding so that algorithms can handle them. Normalization was used to scale numerical values to the same range across features and to minimize the effects of scale differences in different units on model learning, especially for distance-based models, such as KNN. Once pre-processing was completed, meaningful features affecting credit decisions were obtained from a statistically dimensioned model selection method. A variety of machine learning algorithms were then trained, including logistic regression (LR), decision trees (DT), random forests (RF), k-nearest neighbor (KNN), and support vector machine (SVM). These models were comprehensively evaluated using multiple metrics, including accuracy, F1 coefficient, recall, and ROC-AUC, to ensure performance was assessed from various perspectives. Finally, a comparative analysis of the models' performance was conducted to determine the most efficient and stable model in predicting loan risk. This supports the research objective of providing a reliable predictive system that enhances lending decisions in financial institutions.

Table 1: The attributes of the first dataset

No.	Attribute name	type	Description
1	Id	Integer	
2	Student	String	Yes, No
3	Balance	Numeric	
4	Income	Numeric	
5	Default	String	Yes, No

3.1 Data Collection

To build an accurate and robust predictive model for loan default detection, two publicly available datasets were collected from Kaggle to serve as the foundation for analysis. Both datasets were obtained in CSV format and contain structured tabular data relevant to personal and financial characteristics of loan applicants. These datasets are essential for exploring factors contributing to loan default and training the model for classification and risk prediction tasks. The first dataset[18] includes 10,000 records with attributes such as the loan applicant's balance, income, and student status. The target variable, Default, is binary (Yes/No), indicating whether the applicant defaulted on the loan. Table 1 presents the attributes of this dataset.

Table 2: Attributes of the second dataset

No.	Attribute name	type	Description
1	Loan ID	Numeric	A unique identifier for each loan
2	Age	Numeric	The Age of the borrower
3	Income	Numeric	The annual income of the borrower
4	Loan Amount	Numeric	The amount of money being borrowed
5	Credit Score	Integer	The credit score of a borrower indicates their creditworthiness
6	Months Employed	Integer	The number of months the borrower has been employed
7	Num Credit Lines	Integer	The number of credit lines the borrower has open
8	Interest Rate	Float	The interest rate of the loan
9	Loan Term	Integer	The term length of the loan in months
10	DTI Ratio	Float	The debit income ratio
11	Education	String	High school, Bachelor's, Master's, and PhD
12	Employment Type	String	Unemployed, Full-time, Self-employed, and Part-time
13	Marital Status	String	Single, Married, and Divorced
14	Has Mortgage	String	Yes and No
15	Has Dependents	String	Yes and No
16	Loan Purpose	String	Auto, Other, Business, Home, and Education
17	Has CoSigner	String	Yes and No
18	Default	Integer	0 and 1

The second [19] consists of 10,213 entries and contains richer features, including credit score, employment history, education level, and loan purpose. The Default column here is also binary (0 for no default, 1 for default). As shown in Table 2, this dataset allows for more complex feature engineering and provides detailed information that can enhance the model's predictive performance. These two datasets provide complementary views of loan applicants—one more focused on simplified financial data and the other offering a multidimensional socio-economic

profile. This dual-source data approach supports comprehensive analysis and strengthens the model's generalization ability across different borrower profiles.

Table 3: The Original Values of the First Dataset

id	student	balance	income	default
1	No	729.526495	44361.63	No
2	Yes	817.180407	12106.13	No
3	No	1073.54916	31767.14	No
4	No	529.250605	35704.49	No
5	No	785.655883	38463.5	No
6	Yes	919.588531	7491.559	No
7	No	825.513331	24905.23	No
8	Yes	808.667504	17600.45	No

Pre-processing

The preprocessing phase is critical in preparing datasets for machine learning, ensuring the data is clean, structured, and suitable for analysis. It involves several key sub-steps: first, data cleaning, where both datasets are examined for missing values, duplicate entries, and inconsistent formatting, and irrelevant or redundant data is removed to ensure alignment (e.g., matching IDs between datasets). Next, string values are converted to numeric formats, as most machine learning algorithms require numerical inputs. This step typically involves techniques like label or one-hot encoding to convert categorical features into numbers.

Table 4: The original Values for the Second Dataset

LoanID	Age	Income	Loan Amount	Credit Score	Months Employed	NumCredit Lines
I38PQUQS96	56	85994	50587	520	80	4
HPSK72WA7R	69	50432	124440	458	15	1
C1OZ6DPJ8Y	46	84208	129188	451	26	3
V2KKSFM3UN	32	31713	44799	743	0	3
EY08JDHTZP	60	20437	9139	633	8	4
A9S62RQ7US	25	90298	90448	720	18	2
H8GXPAOS71	38	111188	177025	429	80	1
0HGZQKJ36W	56	126802	155511	531	67	4
1R0N3LGNRJ	36	42053	92357	827	83	1
CM9L1GTT2P	40	132784	228510	480	114	4
IA35XVH6ZO	28	140466	163781	652	94	2
Y8UETC3LSG	28	149227	139759	375	56	3
RM6QSRHIYP	41	23265	63527	829	87	4
GX5YQOGROM	53	117550	95744	395	112	4
X0BVPZLDC0	57	139699	88143	635	112	4
O5DM5MPPNA	41	74064	230883	432	31	2
ZDDRGVTEXS	20	119704	25697	313	49	1
9V0FJW7QPB	39	33015	10889	811	106	2
O1IKKLC69B	19	40718	78515	319	119	2
F7487UU2BF	41	123419	161146	376	65	4
7ASF0IHRIT	61	30142	133714	429	96	1
A22KI1B6SE	47	146113	100621	419	55	1
1MUSHWD9TW	55	132058	130912	583	48	4
LXK7UEMLK0	19	118989	123300	528	73	3
995RE1TIB4	38	56848	168918	468	73	1
D17PDP8LBL	50	81649	78193	839	110	1
C35RYEXWJ0	29	114651	197648	343	58	3
G8AIMX5E52	39	17633	167105	514	62	3
BJNLQ0H95H	61	62519	29676	462	16	1
YIGLFWKNH5	42	141412	197764	580	57	2
GAA8OQN796	66	39568	58945	604	37	4
P3EX8G0AYT	44	100284	225403	551	31	1

Finally, normalization is applied to scale feature values into a standard range, such as [0,1] or [-1,1], ensuring that features with large numeric ranges do not dominate those with smaller ranges, thus improving the performance and convergence of many machine learning models, especially

distance-based ones like KNN. Together, these preprocessing steps ensure the data is clean, consistent, and ready for effective feature extraction and model training, ultimately improving the performance of the machine learning models [20-24]. Two datasets containing the numeric and string values are shown in Tables 3,4, and 5, respectively.

Table 5: The original values in the second dataset (continued)

Interest Rate	Loan Term	DTI Ratio	Education	Employment Type	Marital Status
15.23	36	0.44	Bachelor	Fulltime	Divorced
4.81	60	0.68	Master	Fulltime	Married
21.17	24	0.31	Master	Unemployed	Divorced
7.07	24	0.23	High School	Fulltime	Married
6.51	48	0.73	Bachelor	Unemployed	Divorced
22.72	24	0.1	High School	Unemployed	Single
19.11	12	0.16	Bachelor	Unemployed	Single
8.15	60	0.43	PhD	Fulltime	Married
23.94	48	0.2	Bachelor	Selfemployed	Divorced
9.09	48	0.33	High School	Selfemployed	Married
9.08	48	0.23	High School	Unemployed	Married
5.84	36	0.8	PhD	Fulltime	Divorced
9.73	60	0.45	Master	Fulltime	Divorced
3.58	24	0.73	High School	Unemployed	Single
5.63	48	0.2	Master	Parttime	Divorced
5	60	0.89	Master	Unemployed	Married
9.63	24	0.28	PhD	Unemployed	Single
13.56	60	0.66	Master	Selfemployed	Single
14	24	0.17	Bachelor	Selfemployed	Divorced
16.96	60	0.39	High School	Selfemployed	Single
15.58	12	0.65	PhD	Parttime	Divorced
9.32	12	0.38	Bachelor	Unemployed	Married
5.82	60	0.47	High School	Unemployed	Married
15.29	36	0.22	PhD	Parttime	Single
19.1	24	0.22	Bachelor	Unemployed	Single
21.41	48	0.5	Master	Parttime	Married
21.07	24	0.19	Bachelor	Parttime	Married
7.86	36	0.66	High School	Fulltime	Single
23.91	48	0.12	Bachelor	Unemployed	Divorced
10.18	12	0.19	Bachelor	Fulltime	Married
6.67	12	0.1	High School	Unemployed	Divorced
18.77	36	0.17	Master	Unemployed	Divorced
16.11	60	0.44	Master	Unemployed	Married

The string values in the first dataset are transformed to numeric values, such as student (yes=1 and no=0) and default value (yes=1 and no=0), as shown in Table 6. The second dataset converts values into strings and numeric values, as shown in Table 7. Then, the missing values are estimated and normalised to increase the performance of ML models.

Table 6: The converted values in the first dataset

id	student	balance	income	default
1	0	729.526495	44361.63	0
2	1	817.180407	12106.13	0
3	0	1073.54916	31767.14	0
4	0	529.250605	35704.49	0
5	0	785.655883	38463.5	0
6	1	919.588531	7491.559	0
7	0	825.513331	24905.23	0
8	1	808.667504	17600.45	0
9	0	1161.05785	37468.53	0

Table 7: String attribute values after conversion in the second dataset

No.	Attribute name	Description
1	Education	High school=1, Bachelor=2, Master=3 and PhD =4
2	Employment Type	Un employed=0, Full time=1, Self-time=2 and Part time=3
3	Marital Status	Single=0, Married=1 and Divorced =2
4	Has Mortgage	Yes=1 and No=2
5	Has Dependents	Yes=1 and No=2
6	Loan Purpose	Auto=0, Other=1, Business=2, Home=3 and Education=4
7	Has Co-Signer	Yes=1 and No=2

3.3 Feature Descriptions.

The bank's default payment data are represented in the dataset. The collection contains 10000records of five features as shown in Table 1. The second dataset includes 10213records and 18 attributes as shown in Table 2.

3.2 ML Techniques and Evaluation Metrics

Several pre-processing techniques are applied to the original value for a dataset to produce a well-formed dataset, including cleaning, data integration, data formatting, data normalisation, etc. Then, various classification methods are implemented to determine the accuracy of our predictive model's predictions, including LR, Naïve Bayes, Hoeffing Tree, J48, DT, K-NN, SVM, RF, RT, REB Tree, and decision stamp. These techniques are implemented using the Weka application. The results were impressive. This study used a dichotomous default payment (yes or no) as the dependent variable. The results are then compared using different evaluation metrics, and the best techniques are selected. The evaluation metrics are the confusion matrix (shown in Fig. 2), correct classification, TP, FP, precision, recall, and ROC [25-32].

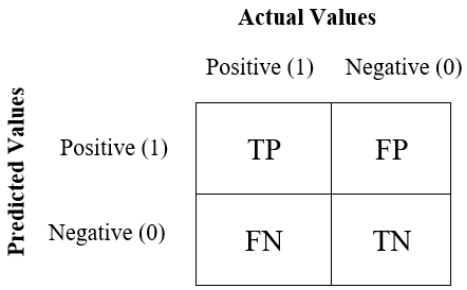

Figure 2: The confusion matrix.

Each prediction will belong to one of the four categories:

- True positive (TP): correct positive prediction
- False positive (FP): incorrect positive prediction
- True negative (TN): correct negative prediction
- False negative (FN): incorrect negative prediction

The data on bank loans is then split into (40%, 50%, 60% and 70%). It is a fundamental and crucial aspect of artificial intelligence. Algorithms are used to analyze data and make relevant decisions. It consists of several branches. In supervised learning, predictions are based on labeled data, and in unsupervised learning, a pattern of unclassified data (patterns) is formed, and reinforcement learning (optimal decisions are made through interactions) [33].

3.2.1 Random Forest (RF)

Fig. 3 shows that Random Forest (RF) is a supervised learning valuable algorithm for regression and classification problems. Model performance is improved through ensemble learning when multiple decision trees (DTs) are standardized. The prediction accuracy is increased and overfitting is avoided, resulting in improved accuracy.[34].

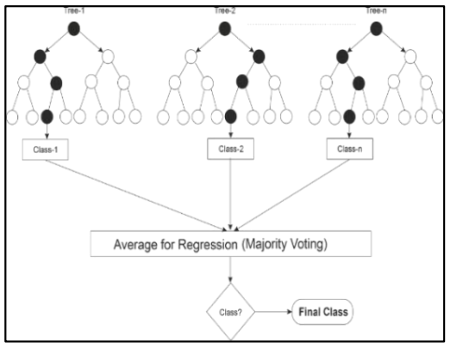

Figure 3: Random Forest Structure [36]

3.2.2 K-Nearest Neighbors (KNN)

KNN algorithm is a popular learning type of ML that can be defined as a supervised learning classifier. An ensemble learning method makes predictions using values from other data points local to the observation [35]. It is used for classification or prediction when grouping individual data points. It is non-parametric because it does not make any underlying assumptions about data distribution. Figure 4 explains two types of data: training data, which classifies coordinates identified by an attribute, and the testing data for identifying the nearest points for the groups query point by determining the closest groups which has the smallest distance, such as Euclidean, Manhattan, and Minkowski distance.

Several studies have implemented a KNN classifier for credit assessment using life credit card data in different types of banks. Other studies have applied and improved different methods to solve the problem of loan defaulters using KNN and LR [35].

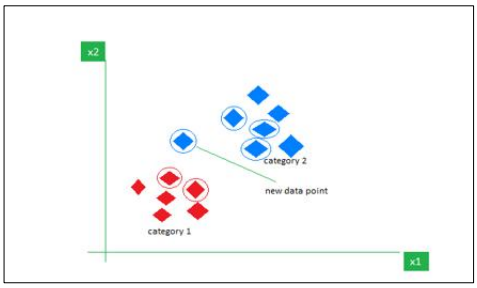

Figure 4: KNN Classifier [37].

3.2.3 Decision Tree (DT)

For classification and regression problems, a supervised learning algorithm called a DT is employed. The tree structure is shown in Fig. 5, where each internal node represents an attribute test. The test's result is a set of branches, and a class label is attached to each leaf node, or terminal node. It is created by iteratively dividing the training data into subsets according to the attribute values until a halting requirement is satisfied, like the maximum tree depth or the smallest number of samples required to split a node. During training, the DT algorithm selects the best attribute to split the data based on a metric such as entropy or Gini impurity, which measures the level of impurity or randomness in the subsets. The goal is to find the attribute that maximises the information gain or the reduction in impurity after the [36]

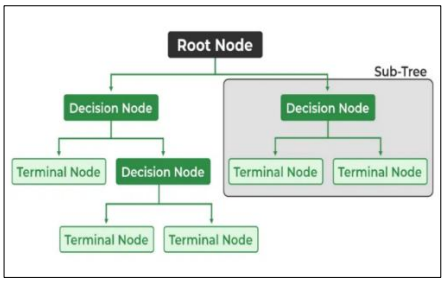

Figure 5: Decision Tree algorithm [38].

3.2.4 Support Vector Machines (SVMs)

Powerful pattern classification models are employed for supervised learning tasks. SVM's effectiveness lies in its ability to handle various learning problems and generalize well to new data. Fig. 6 provides a pseudocode representation of the SVM's process from data splitting to accuracy assessment. [37].

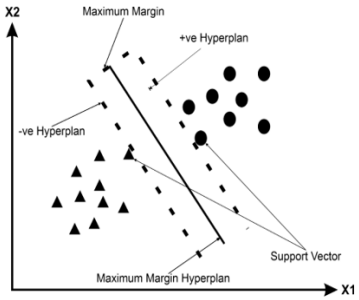

Figure 6: Support Vector Machines (SVMs)[37].

3.2.5 Logistic Regression (LR)

This kind of supervised learning is reliant on discrete or binary variables, such as "yes" or "no," "true" or "false," "spam" or "not spam," etc. A complex cost function known as the logistic function is employed in LR. The following formula is used to calculate this function.

$$f(x) = \frac{1}{1 + e^{x}} \qquad \ldots (1)$$

Where: f(x) is the output of this function between 0 and 1, x is the input value of this function, and e is the natural base.

In the past, LR was used to identify categorization issues in areas including bad debt management, credit scoring, and debt collection. Several studies used machine learning models to increase the speed, effectiveness, and accuracy of the loan approval process. Other research estimated the highest metrics by applying the LR. Still, LR was used in other research, which did not fit well with this kind of data [38].

4. RESULTS AND DISCUSSION

As presented in the section, the main steps for loan prediction consist of five steps as illustrated in Fig. 1, the first step is pre-processing, which is divided into three sub-steps: (i) cleaning data from missing values, (ii) converting the string value to a numeric value in two datasets, and (iii) normalisation. The second step is feature extraction, and the third is ML classification, where different techniques are implemented.

Table 8: Comparison of the Precision of all the Classification Models Performed on the First Dataset

Machine Learning Model	Percent of Training Set			
	40%	50%	60%	70%
Naïve Bayes	97.32	97.5	97.53	97.12
LR	97.45	97.42	97.43	97.10
Hoeffding tree	97.55	91.74	97.18	96.73
Tree-J48	97.42	97.46	97.38	96.90
RF	96.92	97.12	97.33	96.87
RT	95.52	95.44	95.38	95.23
REP tree	97.32	97.56	97.35	97.00
DT	96.77	96.96	97.15	96.80
SVM	96.77	96.96	97.15	96.80

In the fourth step, evaluation metrics are implemented to select the best method using the correct classification rate, TP, FP, precision, recall, and confusion. The pre-processing step and various ML classifiers are implemented on the Weka software. The target of the dataset for classification is a default feature in the first dataset. Then, ML is applied on a training set with different percentages (40%, 50%, 60%, and 70%). Table 8 shows the correct classification results for different ML models using the first dataset.

Table 9: The Correct Classification Results of KNN Based on The First Dataset

Nearest Neighbor	Cross validation		
	15	10	5
1	95.78	95.72	95.67
5	97.02	96.98	97.07
10	97.15	97.11	97.09
15	97.13	97.22	97.17

The top four algorithms are computed using the correct classification rate and are determined to be LR, Hoeffding tree, Tree-J48, and REP tree, with different rates of percentage training set. The highest correct classification rate is 97.56% for the REP tree classifier in 50% of the training set. Moreover, the best algorithm is Naïve Bayes in the 70% training set, computed as 97.53. Other algorithms are calculated at the correct classification as 97.45, 97.55, and 97.42 with 40% training set as LR, Hoeffding tree, and computed 97.53 and 97.12 in 60% and 70% of the training set, respectively, as shown in Fig. 7. KNN is also applied with different numbers of cross validations (10) and nearest neighbours (15) as shown in Table 9. The highest value is 97.22 in 10 cross validations with 15 closest neighbours.

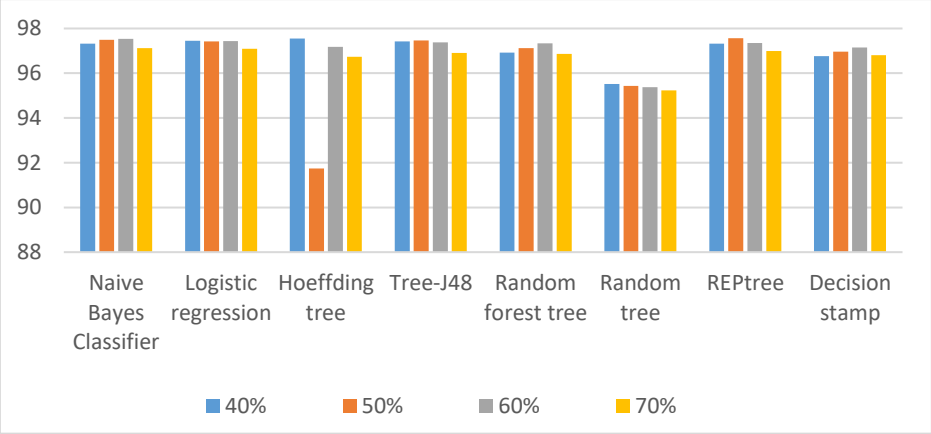

Figure 7: Comparison of the classification rate for all models of the first dataset

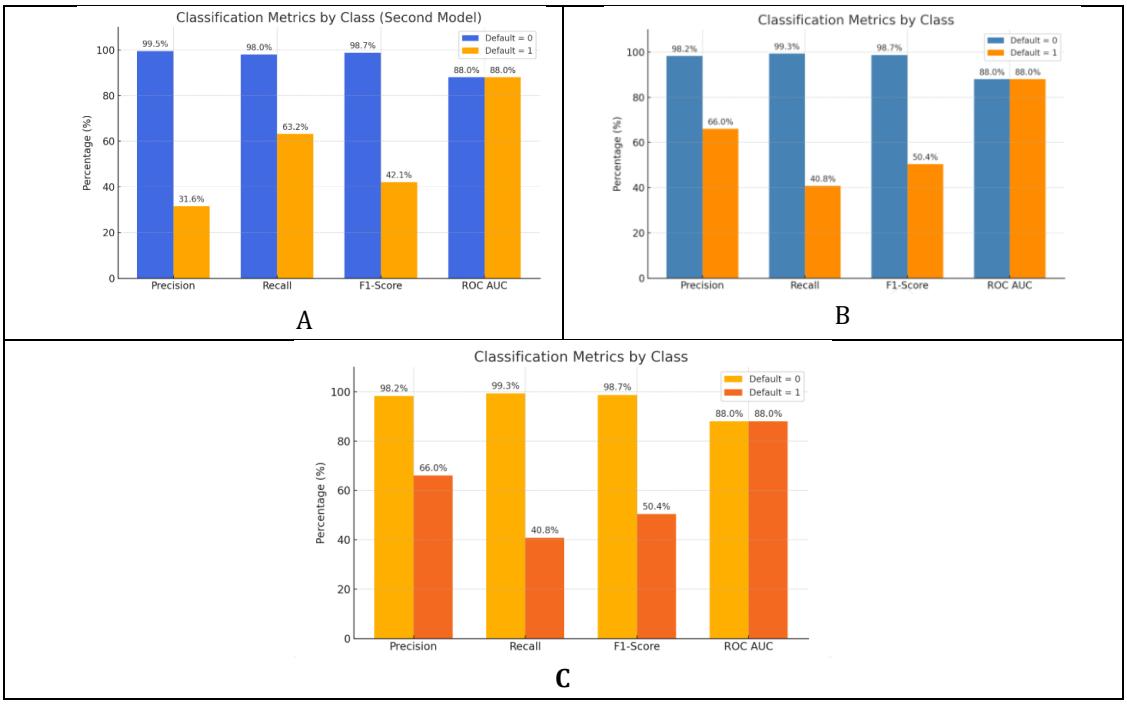

Figure 8: A: Naïve Bayes Classifier with 60% percent of the training set, B: REP tree with 50% percent of the training set, C: The other evaluation metrics for REP tree using the first dataset

Then, several classification metrics are computed to evaluate the best technique, such as the

confusion matrix, correct classification, recall, and precision. The confusion matrix is calculated to select the best ML model. As a result of the bank loan default study utilised in this work, default goal values are binary, i.e., zero represents not having a default, and one represents having one. This evaluation metric is a two-dimensional rectangular array explained in Section 2.7.4. In Table 4.8, the confusion matrix explains the results of the REP tree. Fig. 8 represents the other metrics to evaluate the classification algorithm (REP tree).

Table 10: The Correct Classification Results of Different ML models using the second dataset

Machine Learning Models	Percent of Training 40%	50%	60%	70%
Naïve Bayes Classifier	88.13	87.90	88.00	87.50
LR	88.10	87.88	88.05	87.47
Hoeffding Tree	88.20	87.80	88.03	87.40
J48	85.95	85.94	83.94	83.65
RF	88.04	87.80	88.03	87.47
RT	87.80	87.80	88.03	87.47
REB Tree	88.04	87.80	88.03	87.47
DT	88.04	87.80	88.03	87.47
SVM	88.33	88.29	88.18	88.08

The ML is implemented on the second dataset to compare the results with other features. Table 13 shows the correct classification results with a variety of percentages of training and the five ML models with the highest correct classification. The highest correct classification of SVM with radial basis function is 88.33%, 88.29%, and 88.18%, respectively, with 40%, 50% and 60% rate of training set. The Hoeffding Tree 40% is 88.20, and the Naïve Bayes classifier is 88.13 at the same rate of training set. The lowest is 88.04 for RF, DT, and REB trees, as shown in Fig. 9.

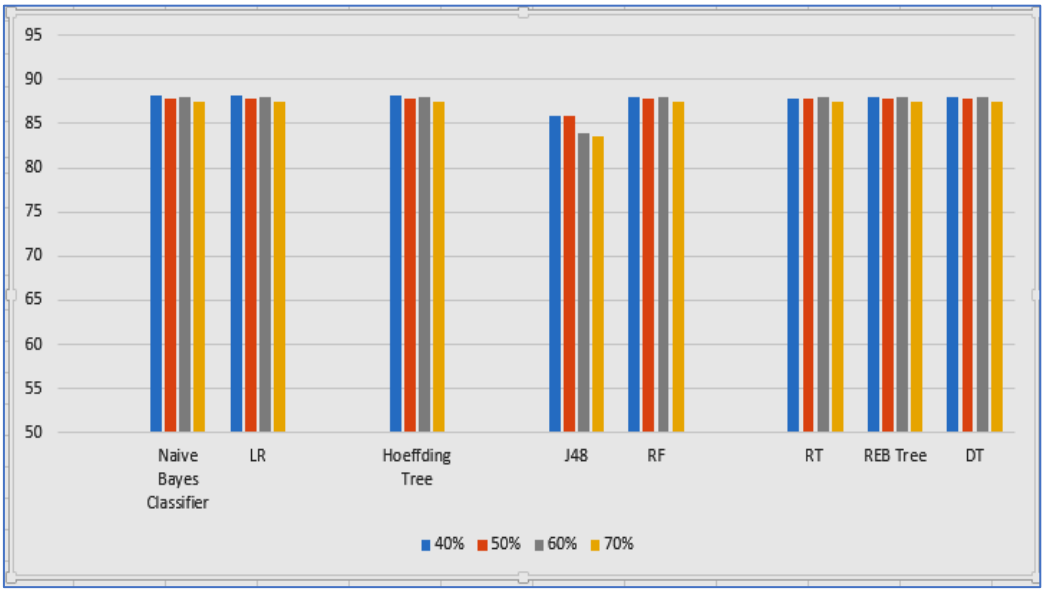

Figure 9: Comparison of Classification Rate of all the ML Models on the Second Dataset.

The KNN is implemented for different cross-validations and nearest neighbours. The highest is 88.15, as shown in Table 11, and Fig. 10 shows the confusion matrix for SVM, and the highest correct classification rate is in the 40% training set. Table 16 is the confusion matrix of SVM in the 50% training set. Tables 17, 18, and 19 show the confusion matrices for Hoeffding Tree, Naïve Bayes (40%), and LR in 40%, respectively.

Table 11: Correct Classification Results of the KNN Algorithm for the Second Dataset

Nearest Neighbor	Cross validation 15	10	5
1	81.39	81.34	81.21
5	87.21	87.26	87.32
10	88.12	88.12	88.14
15	88.14	88.15	88.11

Fig. 10 illustrates the other evaluation metrics for different classification techniques that computed the highest correct classification with varying percentages of the training set.

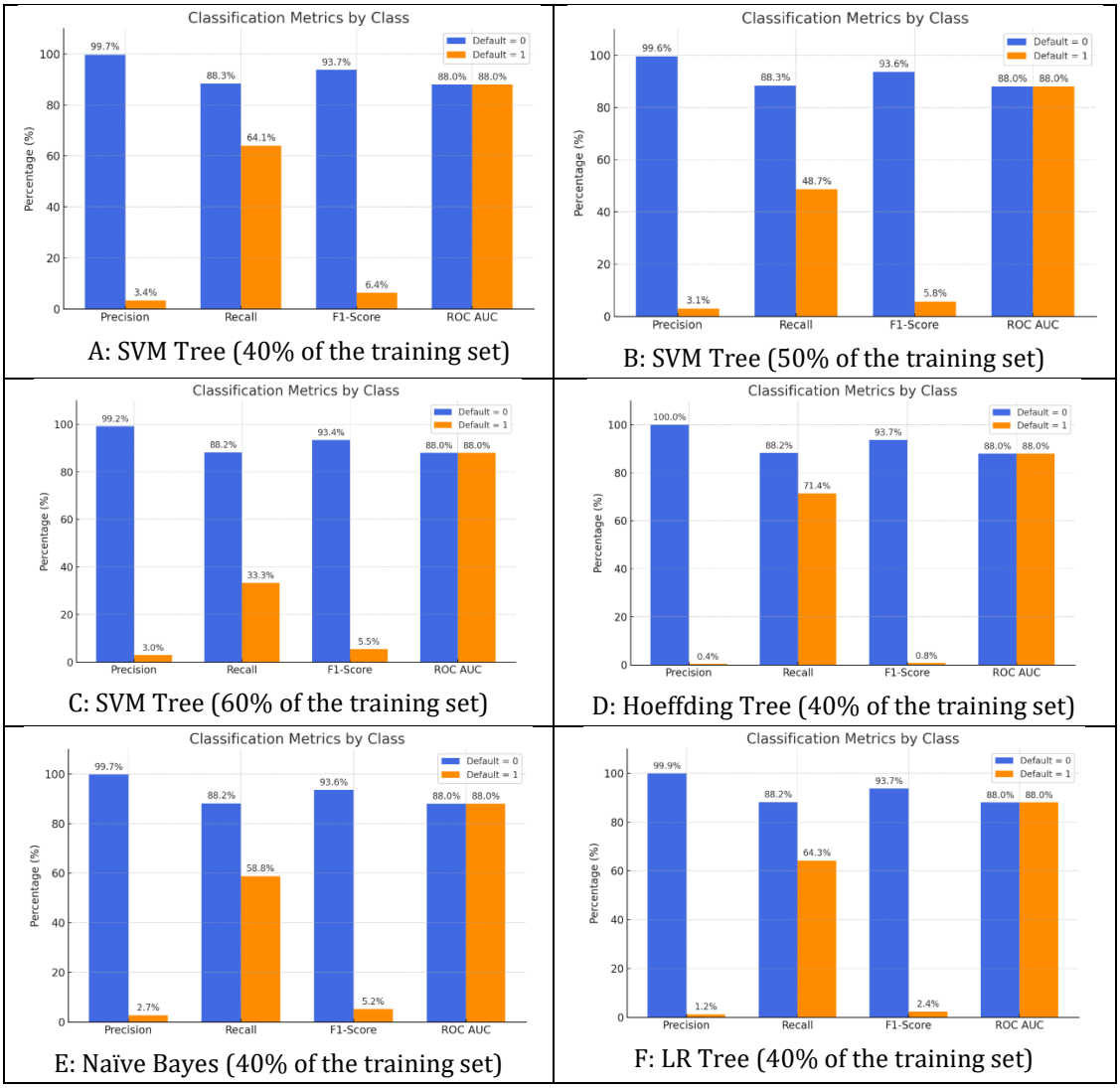

Figure 10: The performance of different models

Table 12: The evaluation metrics for the highest correct classification with different percentage rates of the training set

ML Models		TP	FP	Precision	Recall	F1-Score	ROC Area	Default
SVM	40%			90.00	76.00	82.00		0
				16.00	35.00	22.00		1
	50%			90.00	71.00	79.00		0
				16.00	42.00	23.00		1
	60%			90.00	74.00	81.00		0
				16.00	37.00	22.00		1
Hoeffding Tree		100	99.60	88.20	100	4.80	92.10	0
40%		0.40	0.0	71.40	0.40	4.80	18.10	1
Naïve Bayes		99.70	97.30	88.30	99.70	10.8	95.2	0
40%		2.70	0.30	58.80	2.70	10.8	29.40	1
LR	40%	99.9	98.8	88.20	99.90	93.70	95.10	0
		1.20	0.10	64.30	1.20	2.40	29.10	1
	50%	99.70	98.0	88.20	99.70	8.60	94.90	0
		2.00	0.30	52.60	2.00	8.60	28.70	1

Table 12 shows the results of the correct classification rate of the current work and other researchers implementing different ML models. The results of the thesis work are greater than those of other researchers. R program implemented two ML algorithms, KNN and LR, for two datasets to compare the results with the Weka application.

Table 13: The comparisons of the correct classification results of different ML models from the literature using the second dataset with varying rates of the percentage of the training set

Source	ML Models	Percent of Training				Accuracy (%)
		40%	50%	60%	70%	
Current work	SVM	88.33	88.29	88.18	88.08	
	Naïve Bayes	88.13	87.9	88.0	87.5	
	LR	88.10	87.8	88.0	87.4	
	Hoeffding Tree	88.20	87.8	88.0	87.4	
	J48	85.95	85.9	83.9	83.6	
	RF	88.04	87.8	88.0	87.4	
	RT	87.80	87.8	88.0	87.4	
	REB Tree	88.04	87.8	88.0	87.4	
	DT	88.04	87.8	88.0	87.4	
[6]	RF					91.0
	LR					67.0
	SVM					67.0
[7]	DT					64.0
	RF					77.0

Table 13 shows that the KNN implemented on the first dataset had the highest results on cross-validation 15 and 10, with nearest neighbour 10 computing (97.47). The highest results in KNN with (15) nearest neighbours, with all values in cross-validation as explained in Table 14 and Table 15 on the first and second dataset. Moreover, LR is implemented in two datasets with different percentages of the training set, as shown in Table 16.

Table 14: The classification of KNN based on the first dataset using the R model

Nearest Neighbor	Cross validation		
	15	**10**	**5**
1	96.00	96.00	96.00
5	97.30	97.30	97.30
10	97.47	97.47	97.40
15	97.37	97.37	97.37

Table 15: The classification of KNN based on the second dataset using the R model

Nearest Neighbor	Cross validation		
	15	**10**	**5**
1	81.42	81.42	81.42
5	87.63	87.66	87.63
10	88.02	88.18	88.12
15	88.21	88.21	88.21

Table 16: Comparison of correct classification with (LR) performed on the two datasets

Data sets	Percent of Training Set			
	40%	**50%**	**60%**	**70%**
First data set	96.92	96.34	96.65	96.93
Second data set	88.32	87.68	88.20	87.30

The experimental results demonstrate the notable impact of using different machine learning algorithms and data preparation strategies in loan default prediction. In the first dataset, the REP-Tree classifier achieved the highest classification accuracy of 97.56% at a 50% training ratio, indicating its strong capability to model hierarchical decision structures effectively. Naïve Bayes followed closely, achieving 97.53% at a 70% training ratio, highlighting its ability to generalize well with relatively more training data. The K-Nearest Neighbors (KNN) algorithm also proved reliable, reaching 97.47% using 10 neighbors and 15-fold cross-validation. Logistic Regression (LR) demonstrated stable and consistent results across all training ratios, peaking at 96.93%, which supports its effectiveness in linearly separable datasets. In the second dataset, the Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel outperformed other models, reaching 88.33% accuracy at 40% training, showcasing its strength in handling complex non-linear boundaries. Hoeffding Tree and Naïve Bayes also performed well, achieving classification rates

exceeding 88%. The KNN algorithm showed stable behavior across all configurations, with a peak performance of 88.21% using 15 neighbors. These results collectively affirm that tree-based models such as REP-Tree and Hoeffding Tree, along with instance-based (KNN) and probabilistic (Naïve Bayes) classifiers, are highly effective for predicting loan defaults. Table 17 shows the comparison between the current work and related works.

Table 17: Comparison of the proposed system and related works.

Researcher	Classification Tech.	Results
Current	LR, Hoeffding Tree, J48, REP, NB, KNN, SVM, SVM-RBF, RF, DT	• REP Tree: 98% correct classification rate with 50% training set. (first dataset). • Hoeffding Tree: 88.20% correct classification rate with 40% training set. (second dataset). • KNN: 98% correct classification rate with 15 cross validations and 10 nearest neighbours. (first dataset).
[6]	SVM, LR, and RF	91% RF 67% SVM 67 % LR
[7]	RF and DT	77% RF 68 % DT
[8]	RF and DT	90% RF 87% DT
[9]	ANN- RBF, MLP, and SVM.	84.10% for RBF 78.87% for MLP 76.94% for SVM.
[11]	RF, LG, KNN, and SVM.	RF: 98.04%, LG: 79.60%, KNN: 78.49%, SVM: 68.71%
[12]	RF, KNN, GBoost, DT, SVM, LG	Random Forest: 95.56% ، K Neighbors: 93.33% ، Gradient Boost: 93.33%، Decision Tree: 91.11%، SVM: 84.44%، Logistic Regression: 80.00%
[17]	ANN RBF, LR, ANN-MLP, and SVM	The ANN-RBF was (79.20% for non-defaulters, 97.74% for classifying defaulters, and 75.37% for temporarily defaulters).

5. CONCLUSION

In this work, we tackled the loan default problem by experimenting with different machine learning algorithms and two structured datasets of financial and credit-based attributes. Evaluated models are REP-Tree, Hoeffding Tree, Naïve Bayes, K-Nearest Neighbors (KNN), Logistic Regression

(LR), and Support Vector Machine (SVM), based on varying training test splits. REP-Tree performed the best in dataset 1 (highest accuracy: 97.56%), and SVM (RBF kernel) outperformed other models in dataset 2 (highest accuracy: 88.33%). These results highlight the effectiveness of tree-based models in structured decision problems, the generalization strength of probabilistic methods like Naïve Bayes, and the robustness of KNN under optimal configuration. Based on the findings, model selection should be aligned with dataset characteristics, and training ratios should be carefully tuned to enhance predictive accuracy. These insights can support financial institutions in developing reliable, data-driven systems for credit risk assessment and loan management.

CONFLICT OF INTEREST

The authors declare that there is *no* conflict *of interest* regarding the publication of this paper.

REFERENCES

[1] V. Chang, S. Sivakulasingam, H. Wang, S. T. Wong, M. A. Ganatra, and J. Luo, "Credit risk prediction using machine learning and deep learning: A study on credit card customers," Risks, vol. 12, no. 11, p. 174, 2024, doi: 10.3390/risks12110174.

[2] A. Akinjole, O. Shobayo, J. Popoola, O. Okoyeigbo, and B. Ogunleye, "Ensemble-based machine learning algorithm for loan default risk prediction," Mathematics, vol. 12, no. 21, Nov. 2024, doi: 10.3390/math12213423.

[3] S. Ullah, H. Higgins, B. Braem, B. Latre, C. Blondia, I. Moerman, S. Saleem, Z. Rahman, and K. S. Kwak, "A comprehensive survey of wireless body area networks: On PHY, MAC, and network layers solutions," J. Med. Syst., https://doi.org/10.1007/s10916-010-9571-3.

[4] A. Alagic, N. Zivic, E. Kadusic, D. Hamzic, N. Hadzajlic, M. Dizdarevic, and E. Selmanovic, "Machine learning for an enhanced credit risk analysis: A comparative study of loan approval prediction models integrating mental health data," Mach. Learn. Knowl. Extr., vol. 6, no. 1, pp. 53–77, 2024. https://doi.org/10.3390/make6010004.

[5] A. Alagic, N. Zivic, E. Kadusic, D. Hamzic, N. Hadzajlic, M. Dizdarevic, and E. Selmanovic, "Machine learning for an enhanced credit risk analysis: A comparative study of loan approval prediction models integrating mental health data," Mach. Learn. Knowl. Extr., vol. 6, no. 1, pp. 53–77, 2024. https://doi.org/10.3390/make6010004.

[6] D. Krasovytskyi and A. Stavytskyy, "Predicting mortgage loan defaults using machine learning techniques," Ekonomika, vol. 103, no. 2, pp. 140–160, 2024. https://doi.org/10.15388/Ekon.2024.103.2.8.

[7] R. K. Amin, Indwiarti and Y. Sibaroni, "Implementation of decision tree using C4.5 algorithm in decision making of loan application by debtor (Case study: Bank pasar of Yogyakarta Special Region)," 2015 3rd International Conference on Information and Communication Technology (ICoICT), Nusa Dua, Bali, Indonesia, 2015, pp. 75-80, doi: 10.1109/ICoICT.2015.7231400.

[8] A. Uzair, T. Aziz, H. Ilyas, S. Asim, and B. N. Kadhar, "An empirical study on loan default prediction models," J. Comput. Theor. Nanosci., vol. 16, no. 8, pp. 3483–3488, 2019. https://doi.org/10.1166/jctn.2019.8312.

[9] F. M. Assef and M. T. A. Steiner, "Machine learning techniques in bank credit analysis of companies: A case study of a Brazilian bank," Proceedings of International Conference on Computers and Industrial Engineering, CIE, vol. 2019-October, 2019, Accessed: Jul. 20, 2025. [Online]. Available: https://vbn.aau.dk/en/publications/machine-learning-techniques-in-bank-credit-analysis-of-companies-.

[10] M. A. Sheikh, A. K. Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 490-494, doi: 10.1109/ICESC48915.2020.9155614.

[11] P. S. Saini, A. Bhatnagar and L. Rani, "Loan Approval Prediction using Machine Learning: A Comparative Analysis of Classification Algorithms," 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2023, pp. 1821-1826, doi: 10.1109/ICACITE57410.2023.10182799.

[12] U. E. Orji, C. H. Ugwuishiwu, J. C. N. Nguemaleu, and P. N. Ugwuanyi, "Machine learning models for predicting bank loan eligibility," 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON), Lagos, Nigeria, 2022, pp. 1-5, doi: 10.1109/NIGERCON54645.2022.9803172.

[13] C. Prasanth, R. P. Kumar, A. Rangesh, N. Sasmitha and D. B, "Intelligent Loan Eligibility and Approval System based on Random Forest Algorithm using Machine Learning," 2023 ,International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India, 2023, pp. 84-88, doi: 10.1109/ICIDCA56705.2023.10100225.

[14] S. Dosalwar, K. Kinkar, R. Sannat, and N. Pise, "Analysis of loan availability using machine learning techniques," Int. J. Adv. Res. Sci. Commun. Technol. (IJARSCT), vol. 9, no. 1, pp. 15, Sep. 2021. https://doi.org/10.48175/IJARSCT-189.

[15] N. Robinson and N. Sindhwani, "Loan Default Prediction Using Machine Learning," 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), ICRITO 2024, 2024, doi: 10.1109/ICRITO61523.2024.10522232..

[16] J. A. Gómez, J. Arévalo, R. Paredes, and J. Nin, "End-to-end neural network architecture for fraud scoring in card payments," Pattern Recogn. Lett., vol. 105, pp. 175–181, Apr. 2018. https://doi.org/10.1016/j.patrec.2017.08.024.

[17] X. Li, D. Ergu, D. Zhang, D. Qiu, Y. Cai, and B. Ma, "Prediction of loan default based on multi-model fusion," Procedia Computer Science, vol. 199, pp. 757–764, 2022. https://doi.org/10.1016/j.procs.2022.01.094.

[18] J. Xu, "Factors Influencing Loan Default: An Empirical Analysis Based on Microscopic Evidence," J. Econ. Bus. Manag., vol. 13, no. 1, pp. 1-10, Jan. 2025. doi: 10.18178/joebm.2025.13.1.841

[19] J. C. Cox, D. Kreisman, and S. Dynarski, "Designed to fail: Effects of the default option and information complexity on student loan repayment," J. Public Econ., vol. 192, p. 104298, Dec. 2020, doi: 10.1016/j.jpubeco.2020.104298.

[20] T. M. Alam et al., "An investigation of credit card default prediction in the imbalanced datasets," in IEEE Access, vol. 8, pp. 201173–201198, 2020, doi: 10.1109/ACCESS.2020.3033784.

[21] S. M. Fati, "A loan default prediction model using machine learning and feature engineering," ICIC Express Lett., vol. 18, no. 1, pp. 27–37, 2024. DOI: 10.24507/icicel.18.01.27.

[22] [22] H. Ayari and R. Guetari, "Integrating genetic algorithms and ensemble learning for improved and transparent credit scoring," , Business Information Systems (BIS 2025), K. Węcel, Ed., Lecture Notes in Business Information Processing, vol. 554, Cham: Springer, 2025, https://doi.org/10.1007/978-3-031-94193-1_17.

[23] A. Akinjole, O. Shobayo, J. Popoola, O. Okoyeigbo, and B. Ogunleye, "Ensemble-Based Machine Learning Algorithm for Loan Default Risk Prediction," Mathematics, vol. 12, no. 21, p. 3423, 2024, doi: 10.3390/math12213423.

[24] D. Xu, S. Yuan, L. Zhang and X. Wu, "FairGAN: Fairness-aware Generative Adversarial Networks," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 570-575, doi: 10.1109/BigData.2018.8622525.

[25] S. N. Kalid, K. -C. Khor, K. -H. Ng and G. -K. Tong, "Detecting Frauds and Payment Defaults on Credit Card Data Inherited With Imbalanced Class Distribution and Overlapping Class Problems: A Systematic Review," in IEEE Access, vol. 12, pp. 23636-23652, 2024, doi: 10.1109/ACCESS.2024.3362831.

[26] S. Wattanawongwan, C. Mues, R. Okhrati, T. Choudhry, and M. C. So, "Modelling credit card exposure at default using vine copula quantile regression," Eur. J. Oper. Res., vol. 311, no. 1, pp. 387-399, Nov. 2023, doi: 10.1016/j.ejor.2023.05.016.

[27] M. Khodayari Gharanchaei and P. P. Panda, "Comparison of several machine learning methods in credit card default classification," J. Strateg. Int. Stud., vol. 18, no. 1, pp. 24–31, 2024. https://ssrn.com/abstract=4902470.

[28] A. Subasi and S. Cankurt, "Prediction of default payment of credit card clients using Data Mining Techniques," 2019 International Engineering Conference (IEC), Erbil, Iraq, 2019, pp. 115-120, doi: 10.1109/IEC47844.2019.8950597.

[29] A. Verikas, Z. Kalsyte, M. Bacauskiene, and A. Gelzinis, "Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: A survey," Soft Comput, vol. 14, pp. 995–1010, 2010. https://doi.org/10.1007/s00500-009-0490-5.

[30] J. Ifft, R. Kuhns, and K. Patrick, "Can machine learning improve prediction – an application with farm survey data," Int. Food Agribus. Manag. Rev., vol. 21, no. 8, pp. 1083–1098, 2018.https://doi.org/10.22434/IFAMR2017.0098.

[31] S. K. Saeed and H. Hagras, "A fraud-detection fuzzy logic based system for the Sudanese financial sector," SUST J. Eng. Comput. Sci. (JECS), vol. 20, no. 1, pp. 17–xx, 2019.

[32] A. Kumar, S. Sharma, and M. Mahdavi, "Machine Learning (ML) Technologies for Digital Credit Scoring in Rural Finance: A Literature Review," Risks, vol. 9, no. 11, p. 192, 2021, doi: 10.3390/risks9110192.

[33] T. Xu, "Credit risk assessment using a combined approach of supervised and unsupervised learning," Journal of Computational Methods in Engineering Applications, vol. 4, no. 1, pp. 1–12, 2024. DOI: 10.62836/jcmea.v4i1.040105.

[34] Y. Li, "Credit Risk Prediction Based on Machine Learning Methods," 2019 14th International Conference on Computer Science & Education (ICCSE), Toronto, ON, Canada, 2019, pp. 1011-1013, doi: 10.1109/ICCSE.2019.8845444.

[35] M. Madaan, A. Kumar, C. Keshri, R. Jain, and P. Nagrath, "Loan default prediction using decision trees and random forest: A comparative study," IOP Conf. Ser.: Mater. Sci. Eng., vol. 1022, p. 012042, 2021. [Online]. Available: https://doi.org/10.1088/1757-899X/1022/1/012042.

[36] [38] Aslam, U.; Aziz, T.; Ilyas, H.; Sohail, A.; Batcha, N.; Kadhar, N., "An Empirical Study on Loan Default Prediction Models," Journal of Computational and Theoretical Nanoscience, vol. 16, no. 8, pp. 3483-3488, Aug. 2019, doi: 10.1166/jctn.2019.8312.

[37] Y. Song, Y. Wang, X. Ye, R. Zaretzki, and C. Liu, "Loan default prediction using a credit rating-specific and multi-objective ensemble learning scheme," Information Sciences, vol. 629, pp. 599-617, Jun. 2023, doi: 10.1016/j.ins.2023.02.014.

[38] K. Niu, Z. Zhang, Y. Liu, and R. Li, "Resampling ensemble model based on data distribution for imbalanced credit risk evaluation in P2P lending," Information Sciences, vol. 536, pp. 120-134, Oct. 2020, doi: 10.1016/j.ins.2020.05.040.