Journal of Computer and Communications, 2024, 12, 224-241
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Hybrid 1DCNN-Attention with Enhanced Data Preprocessing for Loan Approval Prediction

Yaru Liu, Huifang Feng*

College of Mathematics and Statistics, Northwest Normal University, Lanzhou, China
Email: *hffeng@nwnu.edu.cn

How to cite this paper: Liu, Y.R. and Feng, H.F. (2024) Hybrid 1DCNN-Attention with Enhanced Data Preprocessing for Loan Approval Prediction. *Journal of Computer and Communications*, **12**, 224-241.
https://doi.org/10.4236/jcc.2024.128014

Received: July 28, 2024
Accepted: August 26, 2024
Published: August 29, 2024

Abstract

In order to reduce the risk of non-performing loans, losses, and improve the loan approval efficiency, it is necessary to establish an intelligent loan risk and approval prediction system. A hybrid deep learning model with 1DCNN-attention network and the enhanced preprocessing techniques is proposed for loan approval prediction. Our proposed model consists of the enhanced data preprocessing and stacking of multiple hybrid modules. Initially, the enhanced data preprocessing techniques using a combination of methods such as standardization, SMOTE oversampling, feature construction, recursive feature elimination (RFE), information value (IV) and principal component analysis (PCA), which not only eliminates the effects of data jitter and non-equilibrium, but also removes redundant features while improving the representation of features. Subsequently, a hybrid module that combines a 1DCNN with an attention mechanism is proposed to extract local and global spatio-temporal features. Finally, the comprehensive experiments conducted validate that the proposed model surpasses state-of-the-art baseline models across various performance metrics, including accuracy, precision, recall, F1 score, and AUC. Our proposed model helps to automate the loan approval process and provides scientific guidance to financial institutions for loan risk control.

Keywords

Loan Approval Prediction, Deep Learning, One-Dimensional Convolutional Neural Network, Attention Mechanism, Data Preprocessing

1. Introduction

Loan approval is an important service in the modern financial system. In today's fast-paced and rapidly evolving society, many individuals and businesses require

additional funds to achieve their dreams and goals. Loan approval as a financial tool is designed to help individuals and businesses get the financial support they need. It is one of the core businesses of financial institutions, but it also involves certain risks.

Today, more and more people rely on loans to meet a variety of consumer needs, including home purchases, education and business startups. As a result, banks encounter a significant volume of loan applications daily. Interest income from loans accounts for a significant portion of bank assets, and lending is an important core business for banks [1]. Therefore, banks place great emphasis on lending to safe and secure borrowers to minimize the risk of default and loss. To ensure this, banks must perform detailed verification and validation of identity and repayment ability of applicants. This includes verifying the applicant's personal information, credit history, income and assets. These validation and verification processes are often cumbersome and require significant time and human resources. Therefore, banks need to establish an effective loan approval system that improves the efficiency and accuracy of approval, reduces risks and losses, and promotes fairness and equality in the loan market [2]. On the one hand, this will provide great convenience to loan approvers, and on the other hand, it can quickly convert some of the deposit users into loan users to expand the loan business. This will have a positive impact on financial institutions, borrowers, and the entire economic system.

With the development of artificial intelligence technology, machine learning models are able to handle massive amounts of data, can effectively extract information from big data, and respond quickly, and have great potential for application in financial decision-making. Prediction in finance mainly includes credit score prediction, loan approval prediction, loan default prediction, etc., which are collectively referred to as loan risk prediction. Machine learning based prediction methods can be broadly categorized into three types, *i.e.*, traditional machine learning models, integrated learning models and deep learning models.

1) **Predictions based on traditional machine learning**: Traditional machine learning methods include support vector machine (SVM), k-nearest neighbor (KNN), naive bayes (NB), logistic regression (LR), decision tree (DT), etc. Arora *et al.* [3] proposed a BootStrap-lasso enabled random forest algorithm (BS-RF) for credit risk assessment. Junior *et al.* [4] proposed the Reduced Minority k-Nearest Neighbors (RMkNN) for imbalance imbalanced credit scoring classification. Fu [5] employed three modelsdecision tree, logistic regression, and support vector machineto forecast loan approval outcomes and evaluated their effectiveness comparatively. The results of the experiments show that the decision trees are the most accurate. Chen [6] introduced a five-level classification approach for agricultural credit utilizing SVM, comparing it against a back-propagation (BP) neural network, the findings indicated that the SVM model outperforms the BP neural network significantly in predicting commercial credit risk. Sheikh *et al.* [7] developed a logistic regression model for predicting the

problem of loan defaulters and calculated different performance metrics. Based on the performance metrics such as sensitivity and specificity, according to the final result, the model exhibited the top performance. Pandey *et al.* [8] employed four machine learning algorithms centered around classification: LR, DT, SVM, and RF, in order to forecast loan approval problem and compare them based on different evaluation metrics. Based on the experimental findings, it is evident that the SVM algorithm attains the highest prediction accuracy and outperforms others in terms of accuracy.

2) **Predictions based on ensemble learning**: Although machine learning models have been widely used in finance, relying on a single classifier is not enough for large-scale banking data in the real world. Therefore, Ensemble learning is being applied to the field of finance, which improves the reliability and accuracy of predictions by using multiple machine learning models. Uddin *et al.* [9] used an integrated voting model including three models, decision tree, logistic regression, and random forest, to predict loan approvals, and the experimental results revealed that the model outperformed the prediction results of a single model. Li *et al.* [10] introduced a loan risk prediction model based on stacking + CNN. The experimental results demonstrated that the proposed prediction model exhibited superior performance in terms of prediction accuracy and recall when compared to both the single model and other comprehensive models. Bhargav *et al.* [11] employed a novel random forest classifier to contrast with machine learning techniques in the context of loan approval prediction. It showed that random forests are more accurate than decision trees in predicting loan approvals. Zhu *et al.* [12] applied logistic regression, decision trees, XGBoost, and LightGBM models to forecast loan defaults. The results demonstrated that both LightGBM and XGBoost exhibited superior predictive performance compared to logistic regression and decision tree models. Meanwhile, they also adopt a model-agnostic local interpretable method to analyze the prediction results in an interpretable way. The findings indicated that the prediction results were influenced by factors including loan term, loan grade, credit rating, and loan amount.

3) **Predictions based on deep learning**: As the intelligent era unfolds and the need for financial data analysis grows, deep learning emerges as the forefront of innovation in the financial domain. Deep learning can help financial institutions better manage and control risks by building accurate risk models through learning from large amounts of historical data. Yang *et al.* [13] introduced Deep Credit, a peer-to-peer group lending system that automatically derives credit risk insights from user activity sequences on a website using a deep learning architecture. Wu *et al.* [14] investigated four different convolutional neural network (CNN) architectures and the effects of weight initialization, stochastic gradient descent, and momentum function on loan approval prediction. Wu *et al.* [15] investigated the use of a deep neural network (DNN) to predict whether a customer will repay on time or not and compared it with traditional learning methods. Experiments showed that DNN outperformed traditional learning

methods. Xiao *et al.* [16] proposed an automated end-to-end deep learning framework called AutoEIS to predict loan defaults. In this framework, a multi-field perceptual expert hybrid (MfMoE) structure was designed for numerical embedding. Extensive experiments showed that AutoEIS had a strong advantage in default prediction, and all metrics were improved over the classical benchmark models.

Although traditional machine learning has multiple applications applications in financial businesses such as stock price prediction, credit rating, risk assessment and fraud detection, there are some shortcomings in traditional machine learning methods. For example, traditional machine learning algorithms are not capable of handling nonlinear, high-dimensional data and are prone to problems such as overfitting and underfitting when faced with large amounts of complex data. Deep learning is a very effective method for handling big data, automatically learning patterns in complex data and extracting data features through supervised or unsupervised learning. To address the problem of loan approval, a hybrid deep learning model with 1DCNN-attention networks for loan approval prediction is proposed. The primary contributions of this paper can be summarized as:

- A combination of the enhanced data preprocessing techniques such as standardization, SMOTE oversampling, feature construction, RFE, IV, and PCA are applied to eliminate the effects of data jitter, non-equilibrium, and to improve the representation of features while removing irrelevant or redundant features.
- A hybrid deep learning model with multiple preprocessing techniques and 1DCNN-attention network is proposed for loan approval prediction. A hybrid module that combines a 1DCNN with an attention mechanism is proposed to extract local and global multilevel spatio-temporal features.
- Our proposed model's effectiveness has been assessed through comprehensive experiments on authentic datasets.

The organization of this paper is as follows: Section 2 elaborates on the proposed models in detail, including dataset, enhanced data preprocessing, and hybrid 1DCNN-attention modules. Section 2 outlines the methodology and introduces the prediction model. Section 3 presents the experiments and analysis. Lastly, Section 4 provides a summary of the paper and delineates avenues for future research.

2. Methodology

In this section, we propose a novel hybrid 1DCNN-Attention model for loan approval prediction. **Figure 1** outlines the fundamental architecture of the proposed model. The core modules include data preprocessing and multiple hybrid modules.

The data preprocessing includes new feature construction, feature selection and dimensionality reduction. Feature construction is a very vital part of feature

engineering, and its major purpose is to create new features with practical significance by combining existing features, thus capturing more information in the data and enhancing the model's ability to understand and predict the data. Feature selection is an important way to improve the performance of machine learning algorithms by choosing the most impactful features from the original set, the dimensionality of the dataset is reduced. Data dimensionality reduction aims to simplify datasets by transforming them from high-dimensional spaces to lower-dimensional representations, preserving their inherent structure as much as possible.

The hybrid module includes 1DCNN and attention mechanism. In the following sub-sections, we briefly describe these feature selection, dimensionality reduction, 1DCNN and attention mechanism which are the core parts of our proposed model.

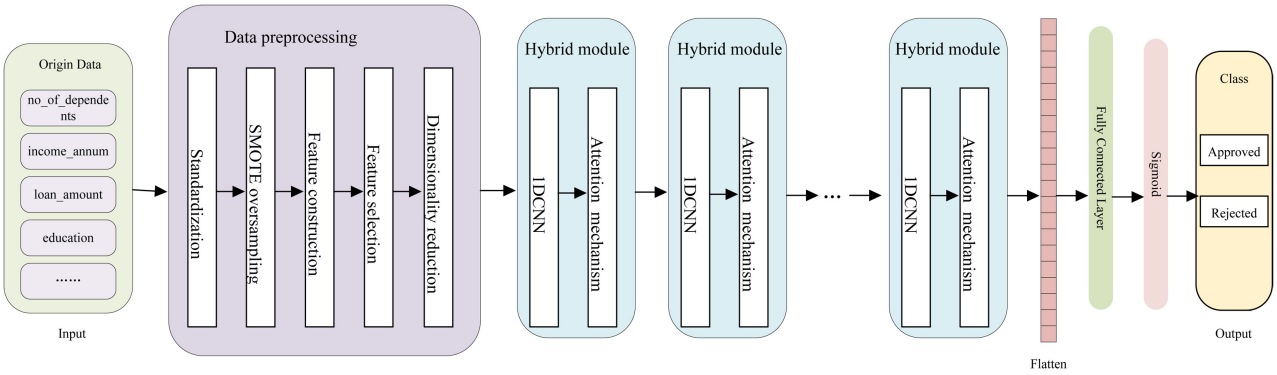

Figure 1. Overview of the proposed model.

2.1. Dataset

The loan approval dataset used in this study is from the kaggle website (https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset). **Table 1** shows the information on the loan dataset. Each sample in the dataset includes 11 features and 1 labeling category. The dataset has a total of 4269 samples, of which 2659 are positive samples and 1613 are negative samples, there is an imbalance of data categories, so the SMOTE method is used for oversampling. In addition, the data need to be standardized in order to eliminate the magnitude differences between features.

2.2. Data Preprocessing

2.2.1. Feature Construction

Feature construction plays a pivotal role in feature engineering, aiming to create new meaningful features by amalgamating existing ones, thus capturing more information in the data and improving the model's ability to understand and predict the data. For example, in this paper, because the loan amount, customer income, asset value and other features only represent the customer's personal situation and the loan situation, which cannot reflect whether the customer has

Table 1. Description of the variables.

Number	Variables	Description
1	no_of_dependents	Number of dependents financially supported by the applicant
2	education	Graduated or not
3	self_employed	Situation of employment
4	income_annum	Annual income of applicants
5	loan_amount	The amount of money requested as a loan
6	loan_term	The duration or term of the loan
7	cibil_score	Credit score of the applicant based on credit information bureau data
8	residential_assets_value	Value of residential assets owned by the applicant
9	commercial_assets_value	Value of commercial assets owned by the applicant
10	luxury_assets_value	Value of luxury assets owned by the applicant
11	bank_asset_value	Value of assets held in bank accounts by the applicant
12	loan_status	The current status of the loan application (e.g., approved, rejected)

the ability to repay the loan, and also can not reflect the repayment risk. Therefore, it is necessary to construct new features based on the existing features, the following new features are constructed and numbered 13-19:

No.13: total_assets_value: it represents the borrower's total assets, total_assets_value = residential_assets_value + commercial_assets_value + luxury_assets_value.

No.14: loan_to_income_ratio: it reflects the weight of the loan relative to the individual's income, loan_to_income_ratio = loan_amount/income_annum.

No.15: total_asset_to_loan_ratio: it reflects the extent to which an individual's assets cover the amount of the loan, total_asset_to_loan_ratio = total_asset_to_loan_ratio/loan_amount.

No.16: total_asset_to_income_ratio: reflects the stability of the individual's financial situation, total_asset_to_income_ratio = total_assets_value/income_annum.

No.17: family_size: reflects the size of the family responsibilities and financial burden borne by the applicant, family_size = no_of_dependents + 1.

No.18: average_annual_income: reflects the applicant's personal economic situation and financial stability, average_annual_income = income_annum/family_size.

No.19: average_annual_loan_amount: it reflects the size of the loan, average_annual_loan_amount = loan_amount/loan_term.

After feature construction, there are 18 features (No.1-11, No.13-19) and 1

label (No.12) in dataset.

2.2.2. Recursive Feature Elimination

Feature selection is the strategic process of identifying the most influential features within the original dataset to minimize its dimensionality, thereby optimizing the performance of machine learning algorithms. In this paper, RFE is combined with IV for feature selection.

Recursive feature elimination (RFE) [17] is a model-based feature selection method, which trains the model iteratively and ranks the features, and then excludes the features with the smallest weights sequentially. The main steps include:

Step 1: Train the model on the training set using all features.

Step 2: Calculate the importance score of each feature.

Step 3: Rank the importance scores of the features and eliminate irrelevant features with low importance scores.

Step 4: Re-train the model on the remaining features and recalculate the importance score.

Step 5: Repeat steps 3 and 4 until the best subset of features is selected.

2.2.3. Information Value

Information value (IV) is an important tool for feature selection in binary classification problems [18]. Suppose a binary target y discretizes a variable x into n bins. The value of Information Value (IV) of predictor x is calculated by the following:

$$\mathrm{IV} = \sum_{i=1}^{k} \left(p_i - q_i \right) \times \ln\left(\frac{p_i}{q_i} \right) \tag{1}$$

where p_i represent the relative frequencies of instances with $y = 1$ at bin i, q_i represent the relative frequencies of instances with $y = 0$ at bin i.

In general, a higher IV tends to indicate stronger predictive power. Table 2 gives the rules for determining the importance of variables (or characteristics) [19].

Table 2. Threshold values for IV.

IV	Predictive Power
<0.02	Not useful for prediction
0.02 to 0.1	Weak
0.1 to 0.3	Medium
0.3 to 0.5	Strong
>0.5	Suspicious Predictive Power

2.2.4. Principal Component Analysis

Principal component analysis (PCA) is a dimensionality reduction method that involves transforming highly correlated features into a small number of inde-

pendent or uncorrelated composite variables through matrix transformations in order to eliminate correlations among the features, thereby reducing the effect of multicollinearity and reducing the dimensionality of the input data. The main steps include:

Step 1: Standardizing data

Using the z-score to standardize the raw data, the formula to compute a Z-score can be expressed as:

$$X_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, i = 1, 2, \cdots, n;\ j = 1, 2, \cdots, p \tag{2}$$

where x_{ij} represents the value of the j-th feature for the i-th instance, μ and σ are denote the mean and the standard deviation for j-th feature, respectively. p and n are denote the number of feature and instance, respectively.

Step 2: Calculating the covariance matrix

The covariance matrix $R = \left(r_{ij}\right)_{p\times p}$ can be calculated by:

$$r_{ij} = \frac{1}{n-1}\sum\left(X_{ki} - \bar{X}_i\right)\left(X_{kj} - \bar{X}_j\right) \tag{3}$$

Step 3: Determining the number of principal components

The eigenvalues $\lambda_j, j = 1, 2, \cdots, p$ and their eigenvectors are obtained from the eigenequation $\left|R - \lambda E\right| = 0$ and arranged in descending order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. The cumulative proportion of variation is $\sum_{k=1}^{i}\lambda_k \Big/ \sum_{k=1}^{p}\lambda_k$, $i = 1, 2, \cdots, p$. The principal components are constructed based on the eigenvectors whose cumulative contribution is greater than a certain threshold.

2.3. Hybrid Module

2.3.1. 1DCNN for the Proposed Model

A one-dimensional convolutional neural network (1DCNN) typically comprises an input layer, convolutional layers, pooling layers, fully connected layers, and an output layer [20]-[22]. The basic structure is shown in **Figure 2**.

1) Input layer: taking a one-dimensional time series (raw input data) and passing it to the first hidden layer.

2) Convolutional layer: convolving the one-dimensional convolutional kernel with the same dimension of the input time series, and extracting the local features by the activation function in the convolutional layer. The convolution process is:

$$x_k^l = \sum_{i=1}^{N_{l-1}} \mathrm{con}\,1D\left(w_{ik}^{l-1}, s_i^{l-1}\right) + b_k^l \tag{4}$$

where s_i^{l-1} is the output of the i-th neurons at layer $l-1$, w_{ik}^{l-1} is the kernel form the i-th neuron at layer $l-1$ to the k-th neuron at layer l, b_k^l is the scalar bias of the k-th neuron at layer l, N_{l-1} represents the number of neurons at layer $l-1$, $\mathrm{con}\,1D(\cdot,\cdot)$ is used to perform the convolutional operator, x_k^l is the input at layer l. The output of the neuron at the hidden layer l, y_k^l, can ob-

tained from the input y_k^l by:

$$y_k^l = f\left(x_k^l\right) \tag{5}$$

where the $f(\cdot)$ is the activation function. The activation functions include Sigmoid, tanh, ReLU, ReLU, etc. In this paper, the rectified linear unit (ReLU) activation function is utilized. Its definition is depicted as follows:

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0 \end{cases} \tag{6}$$

3) Pooling layer: To achieve feature dimensionality reduction and simplify the computational complexity of the network, the feature signals extracted by the convolutional layer are then fed into the pooling layer. The pooling layer is a downsampling operation, it can be illustrated as follows:

$$s_k^l = y_k^l \downarrow ss \tag{7}$$

where $\downarrow ss$ denotes the downsampling operation. The average pooling and the max pooling are the two main downsampling methods.

4) Fully connected layer: Within a fully connected layer, every neuron in the input layer establishes connections with all neurons in the output layer, ensuring comprehensive interconnection between the layers. The fully connected layer produces the output by multiplying the output of the previous layer with the weight matrix, then adding a bias term and performing a nonlinear transformation through the activation function. Therefore, the output in the fully connected layer can be depicted as follows:

$$y_i = f\left(\sum_{j=1}^{M} w_{ij} x_j\right) \tag{8}$$

5) Output layer: The output layer is also known as the classification layer. The output layer of 1DCNN consists of an activation function (e.g., Softmax or Sigmoid) and a loss function. For binary classification problems, a sigmoid activation function is used for calculating the predicted probability for instance, and the binary cross-entropy function is used as a loss function. The sigmoid activation function is defined by Equation (9):

$$\sigma(z) = \frac{1}{1+\mathrm{e}^{-z}} \tag{9}$$

where $z = wx + b$, w and b are the weights and bias between the last two fully connected layers of the neural network, respectively. Then the predicted probability p_j for the j-th instance belongs to the default class is calculated by:

$$p_j = P\left(y_j = 1 \mid x\right) = \sigma(z) = \frac{1}{1+\mathrm{e}^{-(wx+b)}} \tag{10}$$

The binary cross-entropy function is defined by Equation (11):

$$\text{Loss} = -y_j \log\left(p_j\right) + \left(1-y_j\right)\log\left(1-p_j\right) \tag{11}$$

where y_j is the actual label, either 0 or 1, p_j is the predicted probability for the j-th instance. In the backward procedure, the weight values of 1DCNN are

optimized using stochastic gradient descent. Thus, the iterative update formula for the weight values is.

$$w = w - \alpha \frac{\partial \text{Loss}}{\partial w} \tag{12}$$

$$b = b - \alpha \frac{\partial \text{Loss}}{\partial b} \tag{13}$$

where $\frac{\partial \text{Loss}}{\partial w} = x \cdot \sigma(z) - y_j$, $\frac{\partial \text{Loss}}{\partial b} = \left(\sigma(z) - y_j\right)$, α is the learning rate.

2.3.2. Attention Mechanisms

The attention mechanism [23] in neural networks functions as a resource allocation mechanism. It allocates computing resources to prioritize critical tasks while addressing information overload amid constrained computing performance. The core idea behind the attention mechanism lies in selectively focusing limited attention resources on crucial information within a vast dataset, disregarding irrelevant data. This enables the model to highlight and characterize the most critical and relevant information. In this paper, an attention mechanism is introduced into 1DCNN networks, and the attention layer is attached to the end of the convolutional block to improve the situation that the convolutional neural network only focuses on local features, which leads to inaccurate learning of global features.

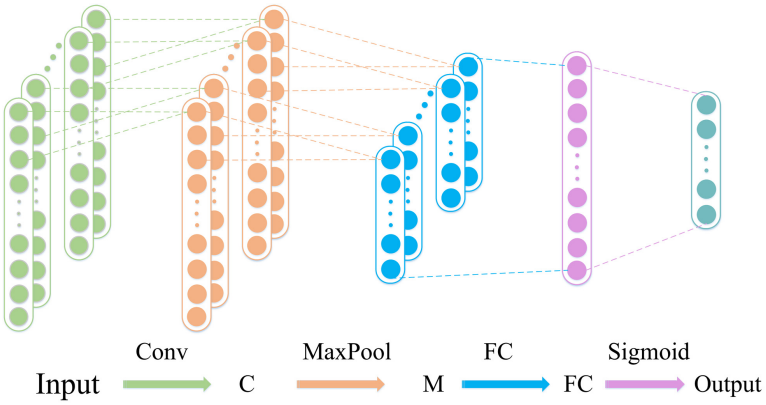

Figure 2. 1DCNN structure.

Let $X = \left[x_1, x_2, \cdots, x_n\right]$ is the output of the 1DCNN networks, the query vector is *q*. Compute the attention distribution α_n:

$$\alpha_n = \text{softmax}\left(s\left(x_n, q\right)\right) = \frac{\exp\left(s\left(x_n, q\right)\right)}{\sum_{j=1}^{N} \exp\left(s\left(x_j, q\right)\right)} \tag{14}$$

where $s(\cdot,\cdot)$ denotes the attention scoring function and can be calculated by the additive model (Equation (15)), dot product model (Equation (16)), scaled dot product model (Equation (17)), or bilinear model (Equation (18)):

$$s\left(x, q\right) = v^{\mathrm{T}} \tanh\left(Wx + Uq\right) \tag{15}$$

$$s(x,q) = x^{\mathrm{T}} q \tag{16}$$

$$s(x,q) = \frac{x^{\mathrm{T}} q}{\sqrt{D}} \tag{17}$$

$$s(x,q) = x^{\mathrm{T}} W q \tag{18}$$

where W, U are learnable parameters and D is the dimension of the input vector. The high-level feature obtained through the attention mechanism is:

$$\alpha = \mathrm{att}(X,q) = \sum_{n=1}^{N} \alpha_n \cdot x_n \tag{19}$$

3. Experimental Results and Analysis

3.1. Parameter Settings and Evaluation Metrics

In this paper, the experiments were executed on a 64-bit Windows 10 platform, and the deep learning frameworks of TensorFlow and Keras are used to construct the network. The experiments are based on the CPU, and the number of cores of the CPU is set to 8 cores.

The benchmark model parameters used in this experiment are Sklearn default parameters and the training set to test set ratio is determined as 8:2. In the model described in this research paper, the training and testing is done using 10-fold cross validation strategy to ensure a good training-testing ratio. The weights for training are optimized using Adam optimizer with Sklearn default parameters. In the experiments, the convolution kernel size is set to 8, the step size is 1, the pooling size is 2, the number of training cycles is 150, and the batch size is 16.

We employ performance metrics such as Accuracy, Precision, Recall, F1 value (F1), and Area Under the ROC Curve (AUC) for evaluation. Equations (20)-(23) define the evaluation metrics described above.

$$\mathrm{Accuracy} = \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{TN} + \mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \tag{20}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{21}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{22}$$

$$\mathrm{F1} = 2 * \frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{23}$$

TP (True positive) represents correctly identified positive samples, while FP (False positive) denotes negative samples incorrectly labeled as positive. TN (True negative) indicates accurately classified negative samples, and FN (False negative) signifies positive samples mistakenly classified as negative.

3.2. Feature Selection and Dimensionality Reduction

The purpose of feature selection and dimensionality reduction excludes these invalid features, simplifies the model and improves the accuracy of prediction. In this paper, RFE is combined with IV for feature selection.

Firstly, the RFE is employed to eliminate features strongly correlated with the target variables. The logistic regression classifier is used as the base model of RFE for feature selection. The order of importance scores for each feature is shown in Figure 3. As can be seen from Figure 3, average_annual_loan_amount and income_annum are important features for loan approval prediction. The education, self_employed, and total_asset_to_loan_ratio with scores less than 0.02 are eliminated, and finally, 15 valid features such as income_annum, cibil_score, and loan_to_income_ratio are retained.

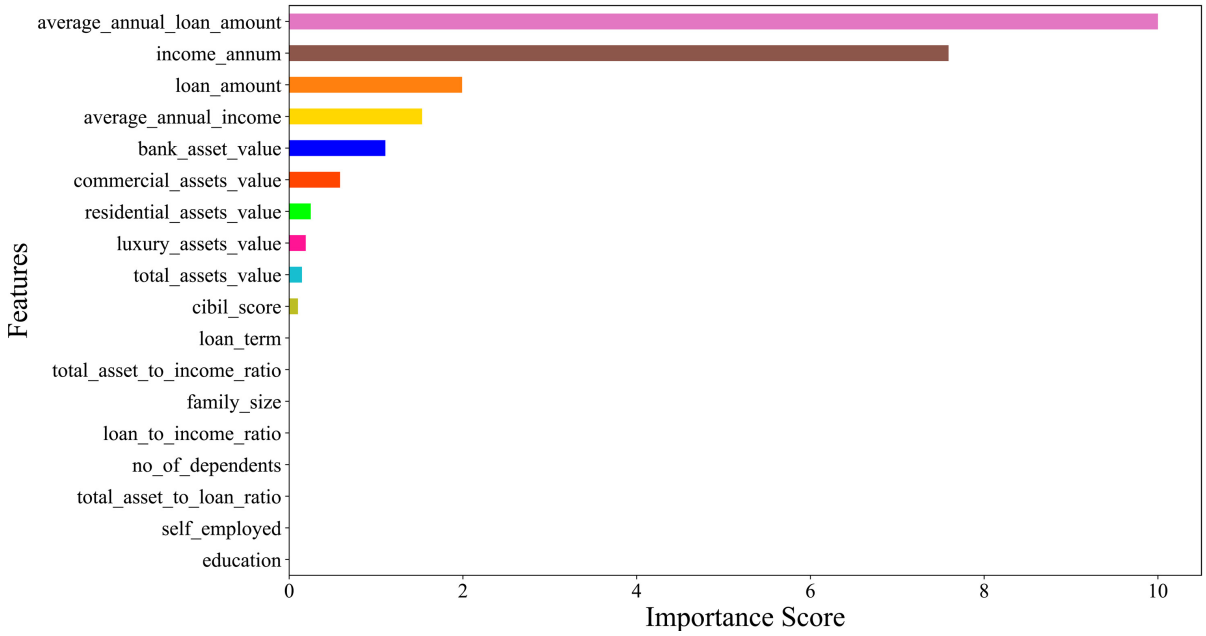

Figure 3. The importance scores of the features.

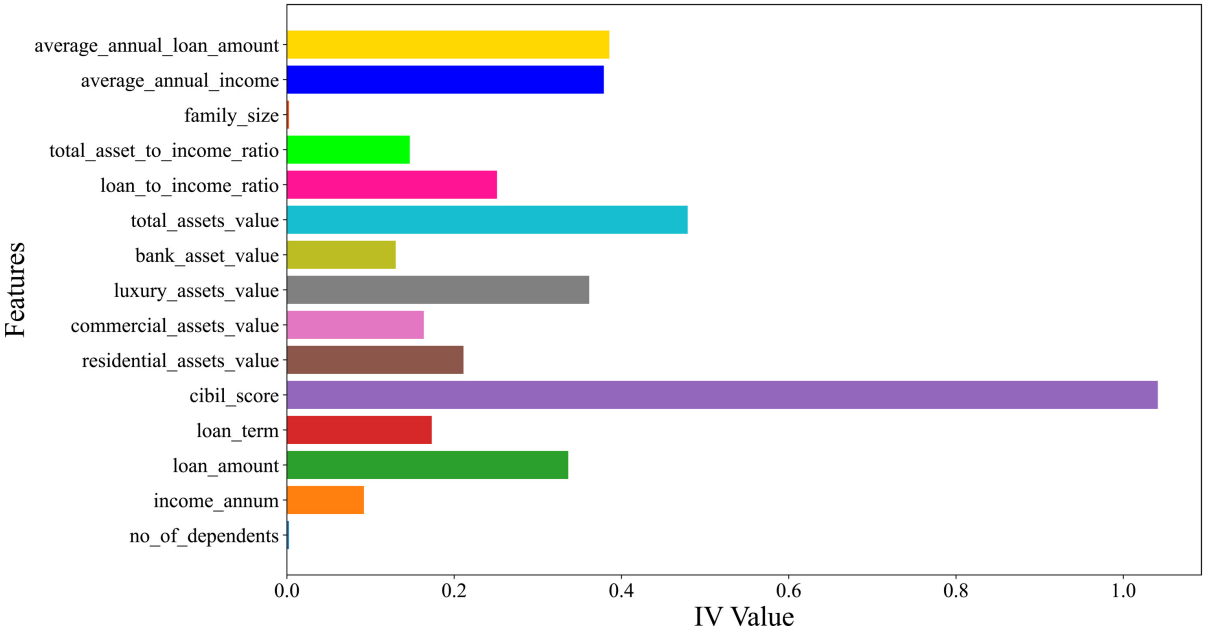

Figure 4. The IV value of the features.

Secondly, the feature selection is then performed using the IV method based on the results of the REF. The findings are illustrated in Figure 4. As depicted in Figure 4, the cibil_score has the highest IV value, so it contributes the most to the loan approval prediction, and the values of feature no_of_dependents and family_size are less than 0.02, so they are removed, and finally 13 features remaining.

Finally, the dimensionality reduction is then executed using the PCA based on the results of the IV. In our case, the threshold of the PCA is set 95%, and select the number of components that the cumulative variance is larger than 95%. Therefore, the eight components are determined for loan approval prediction.

3.3. Comparative Experiments

3.3.1. Comparison of Different Models

To emphasize the outstanding performance of our proposed model, the following baselines are selected as references for comparative studies, including SVM [24], KNN [25], Logistic [26], RF [27], Adaboost [28] and XGboost [29]. The specific experiment results of all models are displayed as in Table 3. In Table 3, the proposed model attains the optimal performance levels on all evaluation metrics, with 99.57% accuracy, 99.42% precision, 99.68% recall, 99.55% F1-score and 0.9998 AUC. Compared to the best results of the evaluation metrics of traditional machine learning algorithms, the accuracy, precision, recall, F1 value and AUC are improved by 1.55%, 1.12%, 1.76%, 1.53% and 0.18, respectively. These results indicate that our proposed model outperforms traditional machine learning.

Table 3. Comparison results for different models.

Model	Accuracy	Precision	Recall	F1	AUC
SVM	0.9389	0.9447	0.9322	0.9384	0.9759
KNN	0.9304	0.9525	0.9058	0.9286	0.9829
logistic	0.9276	0.9299	0.9247	0.9273	0.9765
RF	0.9548	0.9671	0.9416	0.9542	0.9897
Adaboost	0.9492	0.9526	0.9454	0.9490	0.9891
XGboost	0.9802	0.9830	0.9774	0.9802	0.9980
Our model	0.9957	0.9942	0.9968	0.9955	0.9998

3.3.2. Comparison of Different Numbers of Hybrid Modules

Our proposed model consists of a stack of multiple hybrid modules. The number of the hybrid modules has different effects on the feature extraction ability of the data, if too few it is not enough to extract the information contained in the data, if too many it tends to cause overfitting and time-consuming. Table 4 illustrates how varying the number of hybrid modules impacts the model's performance. We can see that when the number of hybrid modules is 3, all evaluation metrics are optimal except the precision rate, and the accuracy, recall, F1 value and AUC

reach 99.57%, 99.42%, 99.68%, 99.55% and 0.9998, respectively.

Table 4. Comparison results for different models.

Stacking	Accuracy	Precision	Recall	F1	AUC
1-layer	0.9806	0.9808	0.9799	0.9803	0.9982
2-layer	0.9947	0.9941	0.9950	0.9945	0.9997
3-layer	0.9957	0.9942	0.9968	0.9955	0.9998
4-layer	0.9955	0.9961	0.9945	0.9953	0.9996
5-layer	0.9921	0.9915	0.9924	0.9919	0.9961

3.3.3. Comparison of Different Data Preprocessing

Data preprocessing is the process of cleaning and transforming raw data to make it more suitable for data analysis and modeling, thereby reducing errors and biases and improving the quality and reliability of the data. The data preprocessing in this paper includes: normalization, SMOTE oversampling, feature construction, RFE, IV, and PCA. The above preprocessing can eliminate the effect of magnitude, balance the dataset, and improve the expression of features while removing irrelevant or redundant features. Table 5 presents a comparison of the performance results for the proposed model using various data preprocessing techniques. It can be seen that the model without any data preprocessing has the worst performance with 69.78% accuracy, 72.87% precision, 82.89% recall, 77.42% F1 value, and 0.7195 AUC. After data standardization the model's accuracy, precision, recall, F1 value, and AUC reached 98.78%, 99.13%, 98.90%, 99.01% and 0.9981, respectively. This result shows that data standardization has a major impact on model performance. The visualization results depicted in Figure 5 demonstrate a steady improvement in model performance with the refinement of data preprocessing techniques. Thus, data preprocessing has a significant effect on the performance of the model.

Table 5. Comparison results of the data preprocessing.

Std	SMOTE	FC	RFE	IV	PCA	Accuracy	Precision	Recall	F1	AUC
						0.6978	0.7287	0.8289	0.7742	0.7195
√						0.9878	0.9913	0.9890	0.9901	0.9981
√	√					0.9912	0.9907	0.9911	0.9908	0.9992
√	√	√				0.9942	0.9925	0.9952	0.9939	0.9995
√	√	√	√			0.9944	0.9932	0.9948	0.9940	0.9996
√	√	√	√	√		0.9953	0.9944	0.9958	0.9951	0.9997
√	√	√	√	√	√	0.9957	0.9942	0.9968	0.9955	0.9998

3.3.4. Comparison of Different Loss Functions

The loss function measures the difference between predicted results and real labels, serving as a crucial component in neural network training. During the

model training phase, the back-propagation algorithm calculates the gradient of the loss function concerning the model parameters. Subsequently, an optimization algorithm, such as gradient descent, is employed to adjust the model parameters, with the objective of minimizing the loss function. Minimizing the loss function enhances both the accuracy and generalization capability of the model, fostering improved performance across various tasks and datasets. Different loss functions are suitable for different problems, and a fitting loss function can be chosen based on the distinctive features and demands of the problem at hand. In this paper, the proposed model is trained using Hinge loss function (HL), cross-entropy loss function (BCE), and exponential loss function (EL), respectively. The experimental outcomes, depicted in Figure 6, reveal that the model's performance is the poorest when trained with the HL loss, while the cross-entropy loss function yields the most favorable results. Hence, the model proposed in this paper adopts the cross-entropy loss function.

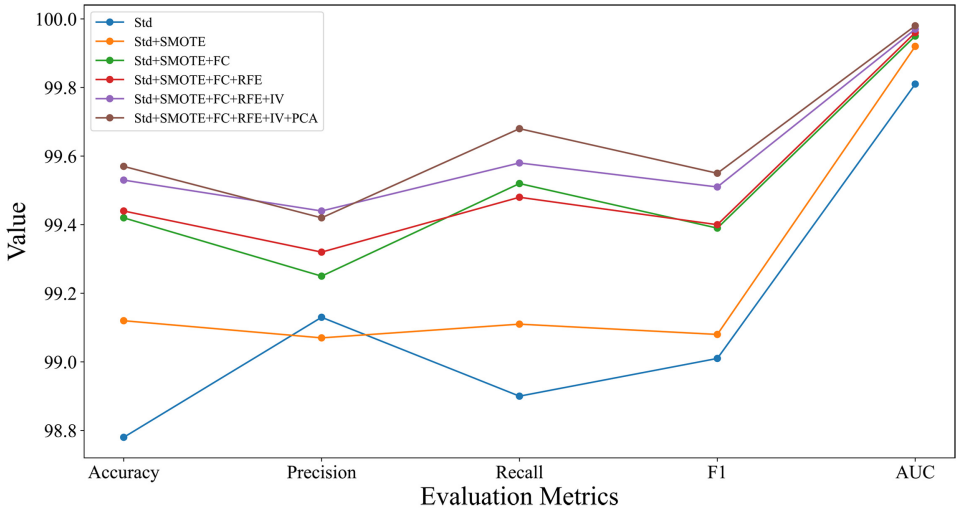

Figure 5. The visualization results for different data preprocessing.

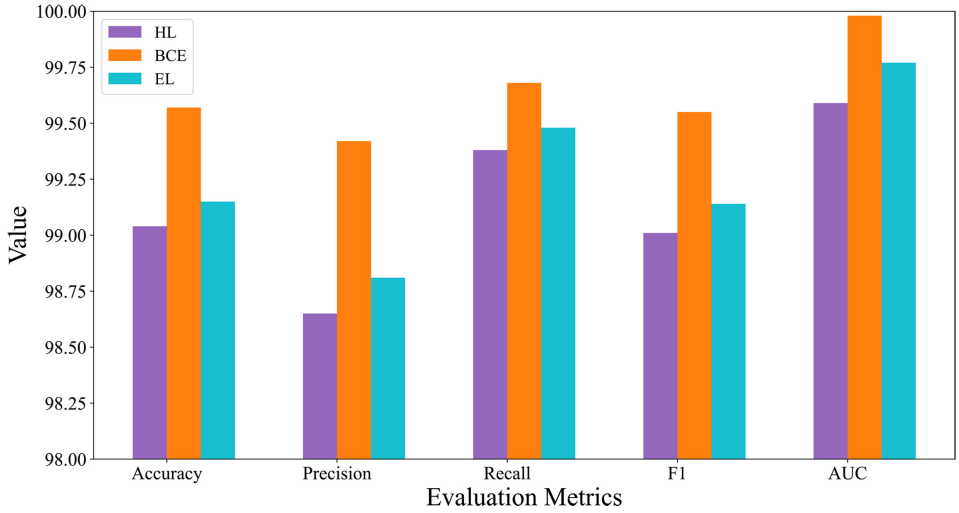

Figure 6. Comparison results of different loss functions.

4. Conclusions and Future Work

The manual operation of loan approval is very cumbersome, and even after rigorous validation and verification, loan approvers are still unable to fully predict and determine which validated applicants are more trustworthy than others, and the risk of borrowing always exists. In order to reduce the risk of non-performing loans, losses, and improve the loan approval efficiency, it is necessary to establish an intelligent loan risk and approval prediction system. A hybrid deep learning model with multiple preprocessing techniques and 1DCNN-attention network is proposed for loan approval prediction. Firstly, a combination of dataset preprocessing techniques is applied to eliminate the effects of data jitter, non-equilibrium, and to improve the representation of features while removing irrelevant or redundant features. Among them, SMOTE deals with data unbalance, RFE and IV are used for feature selection, and PCA is used for dimensionality parsimony. Secondly, a hybrid module that combines a 1DCNN with an attention mechanism is proposed to extract local and global spatio-temporal features. Finally, The comprehensive experiments confirm that our proposed model achieves superior performance compared to state-of-the-art baseline models across multiple metrics including accuracy, precision, recall, F1 score, and AUC. The efficacy of the proposed approach is evident, offering a promising avenue for enhancing online loan risk prediction methods in the future.

The model proposed in this paper achieves good performance on the loan approval prediction dataset, but still has many limitations. First, only one dataset is used in this paper, and due to dataset-specific biases or limitations, our model may not perform as expected on other datasets. Second, we did not perform hyperparameter optimization on our model, and the final model performance may not reach the optimal level. Finally, the process of performing feature construction and feature selection may lead to information loss, which may reduce the model performance. Therefore, in future work, we can conduct experiments using multiple datasets to evaluate the generalization ability and stability of the model. Hyperparametric optimization methods (e.g., grid search, stochastic search, Bayesian optimization, etc.) are used to adjust the model parameters to achieve the optimal level of model performance. More complex feature engineering techniques are used to retain more information during feature construction and selection. Additionally, the development of more effective model interpretation techniques to improve the transparency and understandability of the decision-making process, as well as the application of the models in this paper to other areas of risk management, will be our future work.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Dansana, D., Patro, S.G.K., Mishra, B.K., Prasad, V., Razak, A. and Wodajo, A.W.

(2023) Analyzing the Impact of Loan Features on Bank Loan Prediction Using Random Forest Algorithm. *Engineering Reports*, **6**, e12707. https://doi.org/10.1002/eng2.12707

[2] Sathish Kumar, L., Pandimurugan, V., Usha, D., Nageswara Guptha, M. and Hema, M.S. (2022) Random Forest Tree Classification Algorithm for Predicating Loan. *Materials Today*: *Proceedings*, **57**, 2216-2222. https://doi.org/10.1016/j.matpr.2021.12.322

[3] Arora, N. and Kaur, P.D. (2020) A Bolasso Based Consistent Feature Selection Enabled Random Forest Classification Algorithm: An Application to Credit Risk Assessment. *Applied Soft Computing*, **86**, Article 105936. https://doi.org/10.1016/j.asoc.2019.105936

[4] Melo Junior, L., Nardini, F.M., Renso, C., Trani, R. and Macedo, J.A. (2020) A Novel Approach to Define the Local Region of Dynamic Selection Techniques in Imbalanced Credit Scoring Problems. *Expert Systems with Applications*, **152**, Article 113351. https://doi.org/10.1016/j.eswa.2020.113351

[5] Fu, Y. (2016) A User Loan Approval Evaluation Model and Empirical Study Based on Decision Tree and Support Vector Machine Algorithms. Master's Dissertation, University of Fujian.

[6] Chen, Q. (2020) Research on Rural Commercial Credit Risk Prediction Based on SVM Method. Master's Dissertation, University of Hunan Agricultural.

[7] Sheikh, M.A., Goel, A.K. and Kumar, T. (2020) An Approach for Prediction of Loan Approval Using Machine Learning Algorithm. 2020 *International Conference on Electronics and Sustainable Communication Systems*, Coimbatore, 2-4 July 2020, 490-494. https://doi.org/10.1109/icesc48915.2020.9155614

[8] Pandey, N., Gupta, R. and Uniyal, S. (2021) Loan Approval Prediction Using Machine Learning Algorithms Approach. *International Journal of Innovative Research in Technology*, **8**, 898-902.

[9] Uddin, N., Uddin Ahamed, M.K., Uddin, M.A., Islam, M.M., Talukder, M.A. and Aryal, S. (2023) An Ensemble Machine Learning Based Bank Loan Approval Predictions System with a Smart Application. *International Journal of Cognitive Computing in Engineering*, **4**, 327-339. https://doi.org/10.1016/j.ijcce.2023.09.001

[10] Li, M., Yan, C. and Liu, W. (2021) The Network Loan Risk Prediction Model Based on Convolutional Neural Network and Stacking Fusion Model. *Applied Soft Computing*, **113**, Article 107961. https://doi.org/10.1016/j.asoc.2021.107961

[11] Bhargav, P. and Sashirekha, K. (2023) A Machine Learning Method for Predicting Loan Approval by Comparing the Random Forest and Decision Tree Algorithms. *Journal of Survey in Fisheries Sciences*, **10**, 1803-1813.

[12] Zhu, X., Chu, Q., Song, X., Hu, P. and Peng, L. (2023) Explainable Prediction of Loan Default Based on Machine Learning Models. *Data Science and Management*, **6**, 123-133. https://doi.org/10.1016/j.dsm.2023.04.003

[13] Yang, Z., Zhang, Y.S, Guo, B.H., Zhao, B.Y. and Dai, Y.F. (2018) Deepcredit: Exploiting User Cickstream for Loan Risk Prediction in P2P Lending. *Proceedings of the International AAAI Conference on Web and Social Media*, Palo Alto, 25-28 June 2018, 444-453. https://doi.org/10.1609/icwsm.v12i1.15001

[14] Wu, M., Du, C., Huang, Y., Cui, X. and Duan, J. (2021) Investigation on Loan Approval Based on Convolutional Neural Network. In: Ghosh, A. and Zhou, L.Z., Eds., *Communications in Computer and Information Science*, Springer International Publishing, 203-216. https://doi.org/10.1007/978-3-030-78615-1_18

[15] Wu, M., Huang, Y. and Duan, J. (2019) Investigations on Classification Methods for

Loan Application Based on Machine Learning. 2019 *International Conference on Machine Learning and Cybernetics*, Kobe, 7-10 July 2019, 1-6. https://doi.org/10.1109/icmlc48188.2019.8949252

[16] Xiao, K., Jiang, X., Hou, P. and Zhu, H. (2024) Autoeis: Automatic Feature Embedding, Interaction and Selection on Default Prediction. *Information Processing & Management*, **61**, Article 103526. https://doi.org/10.1016/j.ipm.2023.103526

[17] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002) Gene Selection for Cancer Classification Using Support Vector Machines. *Machine Learning*, **46**, 389-422. https://doi.org/10.1023/a:1012487302797

[18] Rojas, H., Alvarez, C. and Rojas, N. (2013) Statistical Hypothesis Testing for Information Value.

[19] Siddiqi, N. (2012) Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. Wiley and SAS Business Series.

[20] Kiranyaz, S., Ince, T. and Gabbouj, M. (2016) Real-Time Patient-Specific ECG Classification by 1-D Convolutional Neural Networks. *IEEE Transactions on Biomedical Engineering*, **63**, 664-675. https://doi.org/10.1109/tbme.2015.2468589

[21] Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M. and Inman, D.J. (2021) 1D Convolutional Neural Networks and Applications: A Survey. *Mechanical Systems and Signal Processing*, **151**, Article 107398. https://doi.org/10.1016/j.ymssp.2020.107398

[22] Liu, L. and Si, Y. (2022) 1D Convolutional Neural Networks for Chart Pattern Classification in Financial Time Series. *The Journal of Supercomputing*, **78**, 14191-14214. https://doi.org/10.1007/s11227-022-04431-5

[23] Vaswani, A., Shazeer, N., Parmar, N., *et al.* (2017) Attention is All You Need.

[24] Jia, M., Lai, J., Li, K., Chen, J., Huang, K., Ding, C., *et al.* (2024) Optimizing Prediction Accuracy for Early Recurrent Lumbar Disc Herniation with a Directional Mutation-Guided SVM Model. *Computers in Biology and Medicine*, **173**, Article 108297. https://doi.org/10.1016/j.compbiomed.2024.108297

[25] Prasad, B.V.V.S., Gupta, S., Borah, N., Dineshkumar, R., Lautre, H.K. and Mouleswararao, B. (2023) Predicting Diabetes with Multivariate Analysis an Innovative KNN-Based Classifier Approach. *Preventive Medicine*, **174**, Article 107619. https://doi.org/10.1016/j.ypmed.2023.107619

[26] Barboza, F. and Altman, E. (2024) Predicting Financial Distress in Latin American Companies: A Comparative Analysis of Logistic Regression and Random Forest Models. *The North American Journal of Economics and Finance*, **72**, Article 102158. https://doi.org/10.1016/j.najef.2024.102158

[27] Kim, J.H., Lee, D.H., Mendoza, J.A. and Lee, M. (2024) Applying Machine Learning Random Forest (RF) Method in Predicting the Cement Products with a Co-Processing of Input Materials: Optimizing the Hyperparameters. *Environmental Research*, **248**, Article 118300. https://doi.org/10.1016/j.envres.2024.118300

[28] EL Bilali, A., Taleb, A., Bahlaoui, M.A. and Brouziyne, Y. (2021) An Integrated Approach Based on Gaussian Noises-Based Data Augmentation Method and AdaBoost Model to Predict Faecal Coliforms in Rivers with Small Dataset. *Journal of Hydrology*, **599**, Article 126510. https://doi.org/10.1016/j.jhydrol.2021.126510

[29] Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q. and Niu, X. (2018) Study on a Prediction of P2P Network Loan Default Based on the Machine Learning LightGBM and XGboost Algorithms According to Different High Dimensional Data Cleaning. *Electronic Commerce Research and Applications*, **31**, 24-39. https://doi.org/10.1016/j.elerap.2018.08.002