WILEY | Hindawi

Research Article

Digital Universal Financial Credit Risk Analysis Using Particle Swarm Optimization Algorithm with Structure Decision Tree Learning-Based Evaluation Model

Juanjuan Tian and Li Li [ORCID]

School of Management, Wuhan Donghu University, Wuhan 431202, Hubei, China

Correspondence should be addressed to Li Li; lili@wdu.edu.cn

Predictions of credit risk, model reliability, monitoring, and efficient loan processing are all important factors in decision-making and transparency. Machine learning method is providing these people new hope. However, it is up to banking or nonbanking institutions to determine how they will implement this advanced method in order to decrease human biases in loan decision-making. *Objective.* This paper proposed the novel machine learning-based credit risk analysis in the digital banking evaluation model. The purpose of this research is to compare various ML algorithms in order to develop an accurate model for credit risk assessment utilising data from a genuine credit registry dataset. Aim is to design the classification-based model using particle swarm optimization (PSO) algorithm with structure decision tree learning (SDTL) in predicting credit risk. This system has the potential to improve quality criteria such as dependability, robustness, extensibility, and scalability. Features have been extracted and classified by the proposed PSO_SL model. *Experimental Results.* The data have been collected based on real time for credit analysis. Simulation is carried out in Python and optimal results are obtained in comparative analysis with existing techniques. The accuracy obtained by proposed technique is enhanced and the error rate of the design is minimized.

1. Introduction

Because of the increase of rural markets, credit penetration among farmers is gaining traction, and it is becoming recognised as a growth engine for a rising economy such as India. Access to banking, particularly, for smallholder farmers, poorer households, and specific women farmers, is a major issue that must be addressed [1]. These are a few distinct groups of people who are likely currently unserved or neglected by current banking and financial organisations [2]. Their digital transformation of corporate processes is unavoidable as they implement big data solutions to improve their operations [3]. The term "big data" has been popular in recent years, and it refers to massive amounts of data as well as the technologies for storing and analysing them. Banking industry has a significant amount of data that is growing at an exponential rate, and it is grappling with the issues of managing and analysing it. Banks are using big data as well as data science to boost profits by getting new insights from existing data and improving predictions based on that data [4]. To provide forecasts for a variety of expert systems, including liquidity, risk, customer attrition, fraud detection, and revenue, as well as making educated decisions, as a result, AI- and ML-based technologies may be of great assistance to them in determining their credit score [5].

Many researchers have looked at credit risk assessment and assessment models using a variety of techniques. Zhou et al. and Lowd and Davis [6, 7] suggested a particle swarm optimization algorithm-based financial credit risk assessment technique. The study in [8] provided an xgbfs-based financial credit risk assessment approach that decreases the dimension of the user's credit data, trains the xgboost assessment model, and then analyses the user's credit risk. According to [9], combining supervised and unsupervised

ML methods produces better outcomes than employing only one of them. Traditional ways for forecasting credit risks are not well adapted to assist financial institutions, according to [10], and they require ML methods for forecasting credit risk. They introduced a hybrid ensemble ML method that combined the RS (random subspace) and multiboosting, two classic ML methods. They used the usual probit method as well as several ML methods such as NN and KNN in their work [11]. When compared to traditional credit risk assessment methodologies, they achieved a lower mistake rate using ML methods. Kovvuri and Cheripelli [12] utilized a loan dataset from a commercial bank to test five various ML methods for credit risk assessment, including RF, KNN, NB, DT, and logistic regression. The random forest method outperforms the other algorithms studied, according to their findings. The author in [13] built a new P2P credit risk measurement model based on classic credit risk and NN model which are utilized to produce better findings through simulation. The standard credit risk measuring model was combined with SVM and actual study on loan default in [14]. The borrower's loan amount, current overdue loan amount, loan interest rate, credit rating, debt yield, and other parameters are used as independent variables by the researcher in [15], who builds a LR method for testing and gets good results. Dufour et al. [16] replaced the BP neural network's training algorithm with an enhanced PSO method, combined with a credit assessment index system as well as finally realised using an enhanced PSO-BPNN. The author of [17] looked at the origins of P2P online loan risk and used loan data from Lending Club to create a risk prediction model. To provide a credit risk management technique for domestic P2P enterprises, the world's largest P2P company shows forecast accuracy. Restricted Boltzmann machine (RBM algorithm), a multilayer limited Boltzmann introduced by [18], is a nonlinear dimensionality reduction approach employed in this paper. The RBM-DBN model is a DBN model made up of machines. It learns using a multistep unsupervised NN first and then modifies supervised learning parameters before training discriminant classifier model.

The paper contributions are as follows:

(i) On the basis of the credit registry dataset, analyse ML methods in order to construct a precise model for credit risk assessment.

(ii) We concentrate on credit risk assessment, looking at the impact of various proposed models for detecting business defaults. On the contrary, we investigate the models' stability in relation to a subset of variables chosen by the models.

(iii) We design the classification-based model using particle swarm optimization (pso) algorithm with structure decision tree learning (sdtl) in predicting credit risk.

(iv) The findings reveal central banks' perspective on credit risk analysis, which differs significantly from standard commercial bank credit risk evaluations, which rely on more extensive data per client.

2. Digital Banking Credit Risk Analysis Using Particle Swarm Optimization (PSO) Algorithm with Structure Decision Tree Learning (SDTL)

It is important to precisely categorise financial credit risk level, categorise financial credit risk data using PSO-SDTL, and design the credit risk evaluation model to successfully regulate financial credit risk. For various levels of financial credit risk, targeted control measures must be implemented. The overall architecture is shown by Figure 1.

The PSO is started with a set of random particles (solutions) and then updated generations to find the best solution. First is pbest, which is the best solution (fitness) that has been attained thus far. The best value was attained by each particle in population is known as best global, and it is known as the gbest because it performs well in a demanding, nonconvex, and continuous environment. The best value is the best local and is denoted pbest when a particle takes part of population as a neighbour topology.

PSO algorithm is made up of two equations. Because "k" refers to the current iteration as a head, "$k+1$" denotes the next iteration by

$$x_{i+1}^P = x_i^P + v_{i+1}^P. \tag{1}$$

For each combined node k $k \in \{1, \ldots, 2^{h-1} - 1\}$, its output is given by

$$Y_k = \sum_{j=1}^{2} s_{kj} Y_{2k+j-1}. \tag{2}$$

The updating of particle is implemented by each velocity vi that corresponds to a particle. Because PSO is a swarm method, it looks for feasible solutions in m-dimension space in parallel were represented by

$$x_{ij} = x_{ij} + v_{ij}, \tag{3}$$

where i and j indicate ith and jth particle, respectively, given by

$$\min \text{fit} = \min \sum_{\forall x} (Y - Y^*)^2. \tag{4}$$

Consider the discrete state space model is given by

$$x[k+1] = F(x[k], u[k], w). \tag{5}$$

The preceding equation may be recast as the following nonlinear process equation using the definition of $y[k] = x[k+1]$, $k = 0, \ldots, N-1$ is represented by

$$y[k] = F(x[k], u[k], w), \tag{6}$$

where x and u represent inputs, y represents output, and w represents unknown q-dimensional parameter that must be approximated. State space is defined to estimate the parameters given by
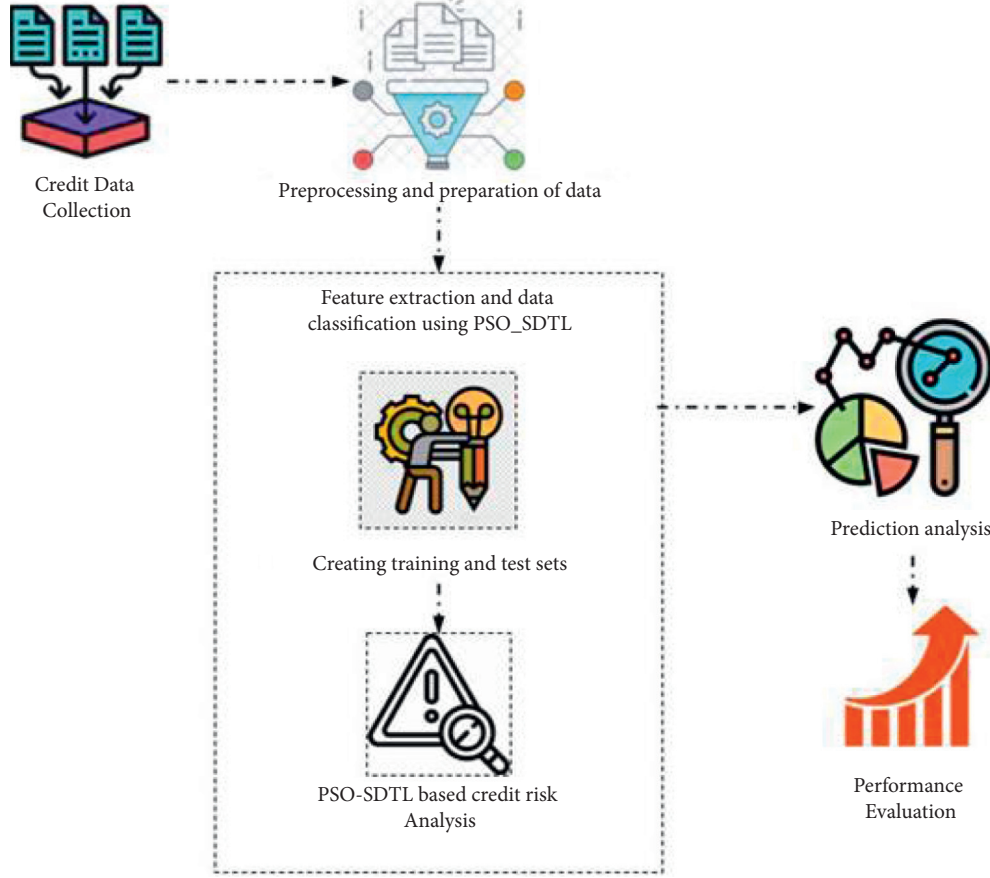
<center>Figure 1: Overall proposed architecture.</center>

$$w[k+1] = w[k] + r[k],$$
$$y[k] = F(x[k], u[k], w[k]) + e[k], \tag{7}$$

where r is the process noise and the first model is the process equation. Latter is measurement equation, which is affected by input as well as measurement noise. Initialize unknown parameter and the covariance matrix were represented by

$$\widehat{\omega}[0] = E[\omega],$$
$$P_{\omega[0]} = E\left[(\omega - \widehat{\omega}[0])(\omega - \widehat{\omega}[0])^T\right]. \tag{8}$$

The following equations provide a set of $2q+1$ sigma vectors W assuming parameter vector has a mean and covariance P_ω given by

$$P_{\omega[k]}^- = P_{\omega[k-1]} + R^r[k-1]W[k|k-1] = \left[\widehat{\omega}^-[k], \widehat{\omega}^-[k] + \gamma\sqrt{P_{\omega[k]}^-}, \widehat{\omega}^-[k] - \gamma\sqrt{P_{\omega[k]}^-}\right]. \tag{9}$$

With nonlinear process F, sigma points are transferred by

$$\mathscr{D}[k|k-1] = F(x[k], u[k], \mathscr{W}[k|k-1]). \tag{10}$$

Transferred sigma point mean and covariance are given by

$$\widehat{d}[k] = \sum_{i=0}^{2q} W_i^{(m)} \mathscr{D}_i[k|k-1], \tag{11}$$

where $W_i^{(m)} = W_i^{(c)} = (1/2(\lambda+q))$ and $W_0^{(m)} = W_0^{(c)} = \lambda/(\lambda+q)$.

The measurement and parameter vectors' cross covariance matrix is computed by

$$P_{\omega,\widehat{d}}[k] = \sum_{i=0}^{2_i} W_i^{(c)} \left[\mathscr{W}_i[k|k-1] - \widehat{\omega}^-[k]\right]\left[D_i[k|k-1] - \widehat{d}[k]\right]^T. \tag{12}$$

In search space, particles are spread based on their own best known position P^* and entire swarm's best known position P_g. In iteration t, velocity of particle i's jth dimension is calculated as follows:

$$P_{ij}(t) = P_{ij}(t-1) + V_{ij}(t). \tag{13}$$

The above description yields the following equation:

$$x_i^{t+1} = x_i^t + v_i^{t+1} \text{ with } x_i^0 \sim U(x_{\min}, x_{\max}),$$

$$P_{\text{best},i}^{t+1} = \begin{cases} P_{\text{best},i}^t, & \text{if } f(x_i^{t+1}) > P_{\text{best},i}^t, \\ x_i^{t+1}, & \text{if } f(x_i^{t+1}) \le P_{\text{best},i}^t, \end{cases} \tag{14}$$

$$G_{\text{best}} = \min\{P_{\text{best},i}^t\}, \quad \text{where } i \in [1,\ldots,n] \text{ and } n > 1.$$

For the gbest PSO method, the velocity of particle I is calculated by

$$\begin{aligned} v_{ij}^{t+1} &= v_{ij}^t + c_1 r_{1j}^t [P_{\text{best},i}^t - x_{ij}^t] + c_2 r_{2j}^t [G_{\text{best}} - x_{ij}^t], \\ v_{ij}^{t+1} &= v_{ij}^t + c_1 r_{1j}^t [P_{\text{best},i}^t - x_{ij}^t] + c_2 r_{2j}^t [L_{\text{best}} - x_{ij}^t]. \end{aligned} \tag{15}$$

The velocity update equation is calculated as

$$\begin{aligned} v_{ij}^{t+1} &= v_{ij}^t, \\ v_{ij}^{t+1} &= v_{ij}^t + c_1 r_{1j}^t [P_{\text{best},i}^t - x_{ij}^t], \\ v_{ij}^{t+1} &= v_{ij}^t + c_2 r_{2j}^t [G_{\text{best}} - x_{ij}^t] \quad \text{for gbest PSO,} \\ v_{ij}^{t+1} &= v_{ij}^t + c_2 r_{2j}^t [L_{\text{best},i} - x_{ij}^t] \quad \text{for gbest PSO.} \end{aligned} \tag{16}$$

When $c_1 = c_2$, all particles are attracted towards the average of $P_{\text{best},i}^t$ and G_{best}.

The analysis of the parameter θ of one dimensional is by

$$dS = \mu(t,\theta,S)dt + \sigma(t,\theta,S)dB(t),$$

$$G(\theta) = g_0(S_0|\theta) \prod_{n=0}^{N-1} g(S_{n+1}|S_n;\theta). \tag{17}$$

Similarly, diminishing log-likelihood function negative effect to calculate θ is shown as

$$S_{n+1} = S_n + \mu(t,\theta,S_n)h + \sigma(t,\theta,S_n)\Delta B_n,$$

$$\frac{1}{\sigma(t,\theta,S_n)\sqrt{2\pi h}} \exp\left[-\frac{(S_{n+1}-S_n-\mu(t,\theta,S_n)h)^2}{2\sigma^2(t,\theta,S_n)h}\right]. \tag{18}$$

The method of moments, on the contrary, is the simplest form on a discrete maximum likelihood function $g(S_{n+1}|S_n;n)$ in the following equation:

$$\overline{\alpha}\left(\overline{\beta}\sum_{n=0}^{N-1} h - \sum_{n=0}^{N-1} S_n h\right) = S_N - S_0,$$

$$\overline{a}\left(\overline{\beta}\sum_{n=0}^{N-1}\frac{h}{S_n} - \sum_{n=0}^{N-1} h\right) = \sum_{n=0}^{N-1}\frac{S_{n+1}-S_n}{S_n}, \tag{19}$$

$$\overline{\sigma}^2 = \frac{1}{N}\sum_{n=0}^{N-1}\frac{(S_{n+1}-S_n-\overline{\alpha}(\overline{\beta}-S_n)h)^2}{S_n h}.$$

Now, we will show you how to learn Markov network structure from data using our method. Each tree is turned into a set of conjunctive characteristics that can reflect the tree's probability distribution.

The primary purpose of organization and regression tree T is to forecast the response variable y as precisely as possible using predictor variables x. Prediction models coupled to leaf nodes define the prediction structure:

$$\mathcal{F} = \left\{(\mathbf{a},\boldsymbol{\mu},\mathbf{C}) \in \mathbb{R}^{p|\tau_B|} \times \mathbb{R}^{|\tau_B|} \times \mathbb{R}^{K|\tau_L|}: \sum_{k=1}^K C_{kt} = 1, \quad C_{kt} \ge 0, \forall k=1,\ldots,K \forall t \in \tau_L\right\},$$

$$P_{it}(\mathbf{a},\boldsymbol{\mu}) = \prod_{t\in\mathcal{N}_t^{\text{left}}} p_{it}(\mathbf{a}\cdot t,\mu_t) \prod_{t\in\mathcal{N}_t^{\text{ight}}} (1-p_{it}(\mathbf{a}\cdot t,\mu_t)), \quad i=1,\ldots,N, t\in\tau_L. \tag{20}$$

For each $k=1,\ldots,K$, the chance of specific $i \in I$ being allocated to class k is equal to $\sum_{t\in\tau_L} P_{it}(\mathbf{a}^*,\mu^*)C_{kt}^*$ by the following equation:

$$\mathbf{x} \longrightarrow \Pi_k(\mathbf{x}) := \sum_{t\in\tau_L} P_{xt}(\mathbf{a}^*,\boldsymbol{\mu}^*)C_{kt}^*, \tag{21}$$

$$\underset{(a,\mu,\tilde{a},\tilde{\mu})\in\mathbb{R}^{(p+1)(^rB)|+|^rL|)}}{\text{mimize}} \sum_{i=1}^N \left(\sum_{t\in\tau_L} P_{it}(\mathbf{a},\boldsymbol{\mu})(\tilde{a}_t^\top x_i + \tilde{\mu}_t) - y_i\right)^2 + \lambda^{\text{local}}\sum_{j=1}^P (a_j,\tilde{a}_j)_1 + \lambda^{\text{global}}\sum_{j=1}^P (a_j,\tilde{a}_j)_\infty. \tag{22}$$

Due to the impossibility of evaluating a closed-form result produced from transitional density (17), a nonparametric kernel density technique was used by the following equation:

$$\overline{g}_M\left[(t, S_n)|(t_{n-1}, S_{n-1}); \theta\right] = \frac{1}{MD}\sum_{i=1}^{M} K\left(\frac{S_n - T_i}{D}\right), \qquad (23)$$

$$g\left[(t_n, S_n)|(t_{n-1}, S_{n-1}), \ldots, (t_0, S_0); \theta\right]. \qquad (24)$$

3. Simulation Results

3.1. Dataset Description

3.1.1. General Data Protection Regulation (GDPR). The dataset serves as the hub for all credit in the country, collecting data from all commercial banks and savings institutions. Each entry in this dataset reflects a monthly credit and credit card status for a certain client. Credit registry database was approximately 1 TB in size, but utilized a subset of this dataset which contains 1,000,000 rows in the study, which signify status of 1,000,000 various credits only for specific clients and their status in planned date of completion of entire loan payment. Dataset lacks a feature that indicates whether or not the client is capable of repaying the loan [19].

3.1.2. Advanced Analytics of Credit Registry Dataset. We showed properties, their dependencies, trends, and dependencies with further data sources using its advanced analytics features. We went through several stages to complete the analysis, including extra calculated columns, measurements, and a star schema model. We developed various reports in a highly efficient manner after model construction and relationships, which, unlike conventional tools, are incomparably quick and powerful with current features for visualisation, dependency, and prediction which allowed us to investigate dataset as well as acquire a better understanding. The speed with which Power BI can manipulate and analyse data was also a factor in our decision. Because Power BI has its own format that is tailored to handle huge data, subset of dataset that evaluated was 13 GB in size when stored in a SQL database, but it was reduced to 330 MB in Power BI format [20, 21].

3.1.3. WIND Dataset. Sample data used in this paper are real data from a commercial bank branch's personal credit database. In this work, 75 data samples from the personal credit database are chosen at random. Following the removal of these invalid samples, 61 valid samples are found. There are 40 learning samples and 21 test samples included in this dataset.

Here, analysis for financial credit risk has been carried out for proposed PSO-SDTL in terms of accuracy, MSE, feasibility, and GDP analysis for various datasets.

Table 1 and Figures 2(a) and 2(b) show performance analysis for proposed PSO_SDTL for GDPR dataset. In Figure 2(a), accuracy and MSE analysis are carried out in which dark blue represents PSO_BPNN and light blue indicates RBM_DBN and 98% and 55% are the accuracy and MSE of the proposed PSO_SDTL method, respectively. In Figure 2(b), feasibility and GDP analysis are done in which dark blue represents PSO_BPNN and light blue indicates RBM_DBN and 98% and 55% are the accuracy and MSE of the proposed PSO_SDTL method, respectively. 94% and 96% are the feasibility and GDP values of proposed PSO_SDTL method, respectively.

Table 2 and Figure 3 show analysis for proposed PSO_SDTL for AACRD dataset. This dataset gives the credit analysis based on analysis of registry dataset. Here, analysing the GDP obtained by proposed PSO_SDTL is 94%, accuracy predicted by this proposed technique is 96% MSE is 52%, and feasibility of analysis is 91%.

Table 3 and Figure 4 show analysis for proposed technique for WIND dataset. This dataset covers the sample credit registry dataset of the bank. The GDP analysis is based on accuracy of the system highly optimized with enhanced feasibility and minimized MSE. By analysing the data, the GDP analysis is optimal for this dataset. The accuracy for this dataset by the proposed model would be 97%.

4. Discussion

Credit risk classification is delicate, and we do not want to miss out on a borrower whose risk goes unnoticed, but it is also critical to ensure that forecasted risk is correct. However, risk assessment for farm assets falls within the binary classification category, and the ML method is best suited for this. Because the feature extraction and preprocessing step is so important for the overall performance of machine learning models, it must be customised for datasets from various fields. Techniques for feature engineering and data mining are constantly evolving. As a result, one of the next projects were excited about is devising a more efficient method of increasing the application's accuracy in making decisions. Another key factor contributing to the models' low accuracy is the small size of the datasets. One dataset contains 1000 entries, whereas the other contains roughly 3000. When high-dimensional datasets have a large number of entries, models perform better. One of the drawbacks of utilising machine learning in credit risk assessment is scarcity of high-quality datasets. When estimating suggested classifiers on unbalanced datasets, such as our credit registry dataset, it is important to keep in mind that the dataset is unbalanced. The results reveal that, on imbalanced data with feature scaling, all of the models perform with good accuracy and precision. Because the results were not promising, we are missing balanced dataset with scaling. Decision tree, random forest, and linear regression are the best models identified by F1 score. The results of feature scaling are poor when data are balanced and scaled results are even worse. When compared to existing articles on datasets from commercial banks, the results obtained by evaluating five machine learning models demonstrate improved outcomes. The findings revealed

TABLE 1: Parametric analysis for GDPR dataset.

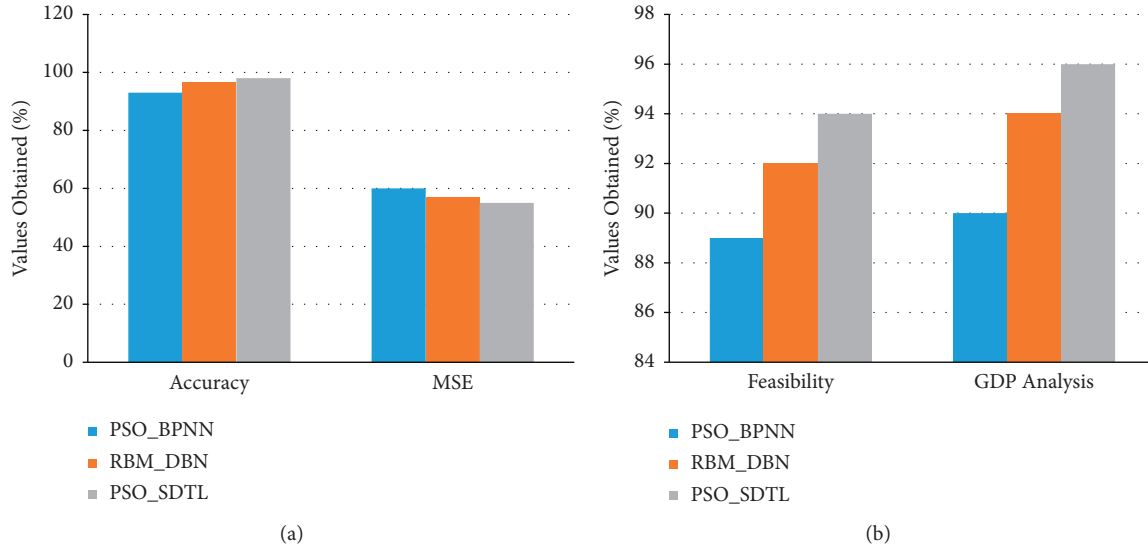|	GDPR dataset		
Specifications	PSO_BPNN	RBM_DBN	PSO_SDTL
Accuracy	93	96.5	98
Mean square error	60	57	55
Feasibility	89	92	94
GDP analysis	90	94	96

FIGURE 2: 2(a) and 2(b) Analysis of GDPR dataset.

TABLE 2: Parametric analysis for the AACRD dataset.

	AACRD dataset		
Specifications	PSO_BPNN	RBM_DBN	PSO_SDTL
Accuracy	91	94	96
Mean square error	57	55	52
Feasibility	85	89	91
GDP analysis	90	92	94

FIGURE 3: Analysis of AACRD dataset.

Table 3: Parametric analysis for WIND dataset.

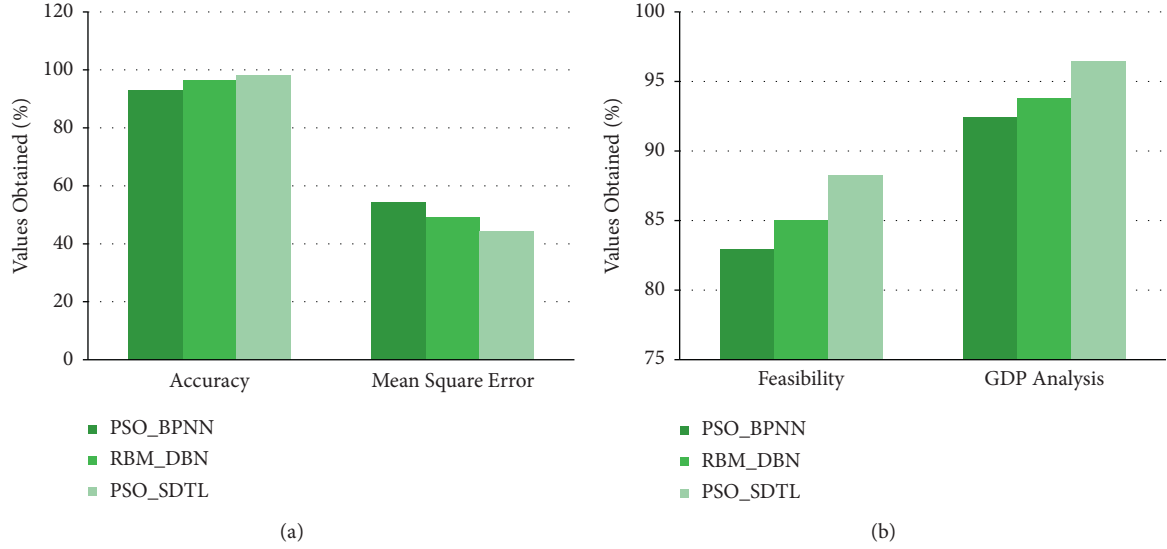| Specifications	WIND dataset		
	PSO_BPNN	RBM_DBN	PSO_SDTL
Accuracy	92	95	97
Mean square error	56	50	45
Feasibility	83	85	88
GDP analysis	92	94	96

(a)

(b)

Figure 4: Analysis of WIND dataset.

```
 (i) Tree Growth (E, F)
 (1) if stopping_cond (E, F) = true then
 (2) leaf = createNode ()
 (3) leaf.label = Categorize (E).
 (4) return leaf.
 (5) Else
 (6) root = generate Node ().
 (7) root.test_cond = determine_best_split (E, F).
 (8) let V = {v|v is a probable outcome of root.test_cond}.
 (9) for each v ∈ V do 10. Eᵥ = {e| root.test_cond (e) = v and e ∈ E}.
(11) child = TreeGrowth (Eᵥ, F).
(12) add child as descendent of root and label edge ( root ⟶ child ) as v.
(13) return root. 14. function LEARNTREE (Xᵢ, D)
(15) best_split ⟵ ∅
(16) best_score ⟵ 0
(17) for Xⱼ ∈ X – Xᵢ do
(18) for xⱼ ∈ Val(Xⱼ) do
(19) S ⟵ (Xⱼ = xⱼ)
(20) if SCORE(S, Xᵢ, D) > best_score then
(21) best_split ⟵ S
(22) best_score ⟵ SCORE(S, Xᵢ, D)
(23) if best_score >logκ then
(24) (Dₜ, D_f) ⟵ SPLITD_ATA (D, best_split )
(25) T_L ⟵ LEARNTREE(Xᵢ, Dₜ)
(26) T_R ⟵ LEARNTREE(Xᵢ, D_f)
(27) return new TreeVertex (b best_split, T_L, T_R)
(28) else
(29) Use D to estimate P(Xᵢ)
(30) return new TrecLeaf (P(Xᵢ))
(31) end if
```

Algorithm 1: Algorithm for SDTL.

that utilising the suggested PSO_SDTL, the models performed well with high accuracy.

5. Conclusion

This paper proposes novel technique in analysis of financial credit risk in digital banking. On the basis of the credit registry dataset, the goal is to test ML methods in order to construct an accurate model for credit risk assessment. We concentrate on credit risk assessment, looking at the impact of various proposed models for detecting business defaults. On the contrary, we investigate the models' stability in relation to a subset of variables chosen by the models. The design of the classification-based model using particle swarm optimization (PSO) algorithm with structure decision tree learning (SDTL) in predicting credit risk has been carried out. The findings reveal central banks' perspective on credit risk analysis, which differs significantly from standard commercial bank credit risk evaluations, which rely on more extensive data per client but lack data for same client in other banks.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgments

References

[1] W. Bao, N. Lianju, and K. Yue, "Integration of unsupervised and supervised machine learning algorithms for credit risk assessment," *Expert Systems with Applications*, vol. 128, pp. 301–315, 2019.

[2] J.-Y. Yeh and C.-H. Chen, "A machine learning approach to predict the success of crowdfunding fintech project," *Journal of Enterprise Information Management*, 2020.

[3] J. Belás, B. G. LubosSmrcka, and D. Jan, "The impact of social and economic factors in the credit risk management of SME," *Technological and Economic Development of Economy*, vol. 24, pp. 1215–1230, 2018.

[4] G. H. Li, X. Yue, W. J. Wu, and J. H. Zhao, "Method of mathematical modeling based on PSO algorithms," in *Applied Mechanics and Materials,* vol. 347-350, pp. 2447–2451, Trans Tech Publications Ltd, 2013.

[5] R. Mosayebi and F. Bahrami, "A modified particle swarm optimization algorithm for parameter estimation of a biological system," *Theoretical Biology & Medical Modelling*, vol. 15, no. 1, pp. 17–10, 2018.

[6] Y. Zhou, S. Liu, T. Tian, Q. He, and X. Ge, "Using particle swarm optimization algorithm to calibrate the term structure model," *Mathematical Problems in Engineering*, vol. 2021, Article ID 8893940, 11 pages, 2021.

[7] D. Lowd and J. Davis, "Learning Markov network structure with decision trees," in *Proceedings of the 2010 IEEE International Conference on Data Mining*, pp. 334–343, IEEE, Sydney, NSW, Australia, December 2010.

[8] D. Lowd and J. Davis, "Improving Markov network structure learning using decision trees," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 501–532, 2014.

[9] E. Carrizosa, C. Molero-Río, and D. Romero Morales, "Mathematical optimization in classification and regression trees," *Top*, vol. 29, no. 1, pp. 5–33, 2021.

[10] F. Doko and I. Miskovski, "Advanced analytics of big data using power BI: credit registry use case," in *Proceedings of the 17th International Conference on Informatics and Information Technologies*, Skopje, North Macedonia, May 2020.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[12] R. S. Kovvuri and R. Cheripelli, "Credit risk valuation using an efficient machine learning algorithm," in *Advances in Decision Sciences, Image Processing, Security and Computer Vision*, pp. 648–657, Springer, Berlin, Germany, 2020.

[13] V. Kuppili, D. Tripathi, D. Reddy Edla, and D. Reddy Edla, "Credit score classification using spiking extreme learning machine," *Computational Intelligence*, vol. 36, no. 2, pp. 402–426, 2020.

[14] K. N. Ohei and R. Brink, "Employability distress: factors affecting ICT graduates' employability and work integration in the workspace," *Research in World Economy*, vol. 12, no. 1, pp. 138–149, 2021.

[15] D. Yang, "Individual credit risk assessment based on stacked," *DenoisingAutoencoder Networks*, vol. 46, no. 10, pp. 7–13, 2019.

[16] B. Dufour, F. Petrella, and N. Richez-Battesti, "Understanding social impact assessment through public value theory: a comparative analysis on work integration social enterprises (WISEs) in France and Denmark," *Working Papers*, vol. 41, no. 27, pp. 112–138, 2020.

[17] Y. Wang, Y. Zhang, Y. Lu, and X. Yu, "A comparative assessment of credit risk model based on machine learning-a case study of bank loan data," *Procedia Computer Science*, vol. 174, pp. 141–149, 2020.

[18] U. P. Shukla and S. J. Nanda, "Designing of a risk assessment model for issuing credit card using parallel social spider algorithm," *Applied Artificial Intelligence*, vol. 33, no. 3, pp. 191–207, 2019.

[19] W. Han, P. Wang, and H. Dong, "Influence of egoistic and altruistic bequest motives on the willingness to participate in reverse mortgages in China," *Asian Economic Journal*, vol. 34, no. 4, pp. 430–463, 2020.

[20] X. Gong, Y. Zhu, H. Zhu, and H. Wei, "Chmusic: traditional chinese music dataset for evaluation of instrument recognition," in *Proceedings of the 2021 4th International Conference on Big Data Technologies*, Zibo, China, September 2021.

[21] H. Song and M. Brandt-Pearce, "A 2-D discrete-time model of physical impairments in wavelength-division multiplexing systems," *Journal of Lightwave Technology*, vol. 30, no. 5, pp. 713–726, 2012.