AIMS Mathematics, 9(2): 4241–4258.
DOI: 10.3934/math.2024209
Received: 15 November 2023
Revised: 30 December 2023
Accepted: 03 January 2024
Published: 16 January 2024

http://www.aimspress.com/journal/Math

Research article

Dandelion optimization based feature selection with machine learning for digital transaction fraud detection

Ebtesam Al-Mansor[1], Mohammed Al-Jabbar[1,*], Arwa Darwish Alzughaibi[2] and Salem Alkhalaf[3]

[1] Computer Sciences Department, Applied College, Najran University, Najran 66462, Saudi Arabia
[2] Applied college, Taibah University, AL Madinah AL Munawwarah, Saudi Arabia
[3] Department of Computer, College of Science and Arts in Ar Rass, Qassim University, Ar Rass, Saudi Arabia

*** Correspondence:** Email: mosalqahtani@nu.edu.sa.

Abstract: Digital transactions relying on credit cards are gradually improving in recent days due to their convenience**.** Due to the tremendous growth of e-services (e.g., mobile payments, e-commerce, and e-finance) and the promotion of credit cards, fraudulent transaction counts are rapidly increasing. Machine learning (ML) is crucial in investigating customer data for detecting and preventing fraud. Conversely, the advent of irrelevant and redundant features in most real-time credit card details reduces the execution of ML techniques. The feature selection (FS) approach's purpose is to detect the most prominent attributes required for developing an effective ML approach, making sure that the classification and computational complexity are improved and decreased, respectively. Therefore, this study presents an evolutionary computing with fuzzy autoencoder based data analytics for credit card fraud detection (ECFAE-CCFD) technique. The purpose of the ECFAE-CCFD technique is to recognize the presence of credit card fraud (CCF) in real time. To achieve this, the ECFAE-CCFD technique performs data normalization in the earlier stage. For selecting features, the ECFAE-CCFD technique applies the dandelion optimization-based feature selection (DO-FS) technique. Moreover, the fuzzy autoencoder (FAE) approach can be exploited for the recognition and classification of CCF. FAE is a category of artificial neural network (ANN) designed for unsupervised learning that leverages fuzzy logic (FL) principles to enhance the representation and reconstruction of input data. An improved billiard optimization algorithm (IBOA) could be implemented for the optimum selection of the parameters based on the FAE algorithm to improve the classification performance. The simulation outcomes of the ECFAE-CCFD algorithm are examined on the benchmark open-access database. The

values display the excellent performance of the ECFAE-CCFD method with respect to various measures.

Keywords: data analytics; evolutionary computation; credit card frauds; machine learning; feature selection
Mathematics Subject Classification: 49-04, 92B20

1. Introduction

Information technology developments have highly influenced the financial industry, resulting in the extensive adoption of electronic commerce (e-commerce) platforms [1]. The main problem related to advanced e-commerce is the optimistic cases of credit card fraud (CCF). In recent years, there has been a growth in CCF that is a great burden on financial organizations. CCF happens in all businesses, ranging from the home appliance to the banking and automotive sectors [2]. Because of the expansive application of credit card fraud detection (CCFD) techniques, users can prevent fraud and be protected from alternative categories of cyber criminals. Automatic fraud detection increases online security and protects users from cybercriminals [3]. Thus, it is important to accurately design automatic fraud detection approaches used for credit card transactions [4]. Several techniques are designed for identifying fraudulent credit card transactions. An increased CCF rate is related to the increasing development of e-commerce and popularity of online transactions. Therefore, CCFD is essential for financial organizations to prevent losses [5].

The machine learning (ML) method has been extensively used for detecting CCF [6]. There are vast databases because of the arrival of the Internet of Things (IoTs) and big data. Due to the size of databases, many features in them may be unrelated or redundant to the response variable [7]. ML can improve the complexity of the model and result in over-fitting by these features. To address the great dimensionality problem, a dimensionality reduction technique like feature selection (FS) is required for obtaining useful insights and making accurate predictions [8]. FS methods aim to detect the most significant features required to design a high-performance ML technique, ensuring decreased computational complexity and enhanced classification performance by extracting redundant and inappropriate features. FS techniques are categorized into three method types: embedded, filter, and wrapper. The internal functioning and configuration of different FS approaches make them suitable for various applications. Filter techniques use feature ranking to determine the useful features. Features that achieve scores more than a given threshold are chosen, and those less than the threshold can be rejected [9]. Subsequently, the identification of key features involves supplying input to the learning method. Filter techniques differ from embedded and wrapper techniques because they are independent of classification bias and are not reliant on the classifier [10,11].

This study presents an evolutionary computing with fuzzy autoencoder based data analytics for credit card fraud detection (ECFAE-CCFD) technique. The ECFAE-CCFD technique performs data normalization in the earlier stage. For selecting features, the ECFAE-CCFD technique applies the dandelion optimization-based feature selection (DO-FS) technique. The global searching abilities of the dandelion optimization (DO) algorithm can efficiently discover the feature space and recognize the highly related features, resulting in significantly better model performance. Moreover, the fuzzy autoencoder (FAE) technique can be implemented for the recognition and classification of CCF.

Autoencoders, particularly FAE, are known for their proficiency in capturing non-linear relationships within data and extracting related features. In CCF, where patterns can be complex and non-linear, FAE can provide effective data representation and improves the classification performance. Last, an improved billiard optimization algorithm (IBOA) can be utilized for the optimum selection of parameters based on the FAE algorithm, increasing the classification accuracy. The IBOA's strategy of escaping local optima can stop the model from getting stuck in suboptimal solutions, guaranteeing improved overall performance. The use of IBOA is motivated by its ability to competently search for optimal parameter values, which is important in improving the performance of the FAE model. The simulation outcomes of the ECFAE-CCFD model are examined on the benchmark open-access database. In short, the contribution of the study will be as follows.

- Introduces the ECFAE-CCFD method, providing an innovative and comprehensive technique for the detection of CCF.
- The DO-FS leverages evolutionary computing to select the most relevant feature for fraud detection, potentially reducing computation complexity and enhancing model performance.
- Employed method represents a significant contribution, which showcases an advanced technique for the detection and classification of CCF. FAE adds a layer of sophistication to the fraud detection model.
- Uses an IBOA for optimum selection of parameters within the FAE algorithm, further increasing the accuracy and generalizability of the model.
- The combination of the DO-FS approach and IBOA for parameter tuning within the FAE framework for the CCFD is an innovative method that has not been discovered in the literature review.

2. Related works

Raghavan and El Gayar [12] target to benchmark multiple ML techniques like support vector machine (SVM), KNN, and RF, while the DL techniques like restricted Boltzmann machine (RBM), autoencoders, CNN, and deep belief network (DBN). These datasets like the German dataset and the European (EU) Australian were used. In [13], current advancements in ML techniques and deep reinforcement learning (DRL) were exploited for CCF detection methods, which include non-fraud and fraud classes. The Adaptive Synthetic Sampling (ADASYN) and Synthetic Minority Over-sampling Technique (SMOTE) were the two resampling approaches leveraged for resampling the imbalanced CCF data. To establish CCF detection, mechanisms like ML techniques were applied to this balanced database. Then, based on the imbalanced CCF database, DRL was used for creating a detection system. Through practical experiments, the author discovered the reliable degree of ML approaches depending on the above-mentioned resampling methods and DRL approaches for the detection of CCF. Alharbi et al. [14] presented the Kaggle dataset to design a DL-related method to sort out the text data problem. The images were given to a CNN structure with class weights through the inverse frequency approach to address the imbalance class problem. ML and DL methods have been implemented to verify the validity and robustness of the presented system.

Sanober et al. [15] introduce a new structure that incorporates Spark with a DL method. This study applies various ML approaches, such as DT, RF, SVM, LR, and KNN, to detect fraud. also, a comparative analysis was done using different parameters. Nguyen et al. [16] offer user separation, where the author splits users into new and old persons, before implementing DNN and CatBoost in all

categories. Also, various methods to boost detection accuracy, like handling feature engineering, heavily imbalanced datasets, and feature transformation, were presented in detail. Almhaithawi et al. [17] addressed fraud detection issues as one common issue in the secure banking research domain because of their significance in decreasing the losses of e-transaction companies and banks. This work includes implementing common classification techniques like LR and RF, along with modern classifiers with existing results such as CatBoost (CB) and XGBoost (XG), testing the outcome of an unbalanced dataset by comparing their outcomes without and with balancing after concentrating on the savings measure for testing the result of cost-sensitive wrapping of Bayes minimum risk (BMR).

Taha and Malebary [18] present an intelligent method to detect CCF transactions utilizing an improved light gradient boosting machine (OLightGBM). A Bayesian-based hyperparameter optimizer method can intelligently combine to adjust the parameters of LightGBM. In [19], the shuffled shepherd political optimizer-based deep residual network (SSPO-based DRN) technique was presented for CCFD. In [20], the authors leveraged the XGBoost method, an effectual method for forecasting appropriately predict fraud. While the important count of fraudulent transactions will be much less than legitimate transactions, the authors offered to set this bias by leveraging sampling approaches like oversampling, undersampling, and ITS combination. Karthik et al. [21] examined a new model for CCFD that integrates ensemble-learning approaches like bagging and boosting. This method combines the main features of both methods by creating a hybrid method of bagging and boosting ensemble methods.

3. The proposed model

In this article, we have proposed the ECFAE-CCFD methodology. The major purpose of the ECFAE-CCFD method is to detect the presence of CCF in real time. To achieve this, the ECFAE-CCFD technique comprises data normalization, IBOA-based parameter tuning, FAE classification, and DO-FS-based feature subset selection. The overall working process of the ECFAE-CCFD technique is depicted in Figure 1.

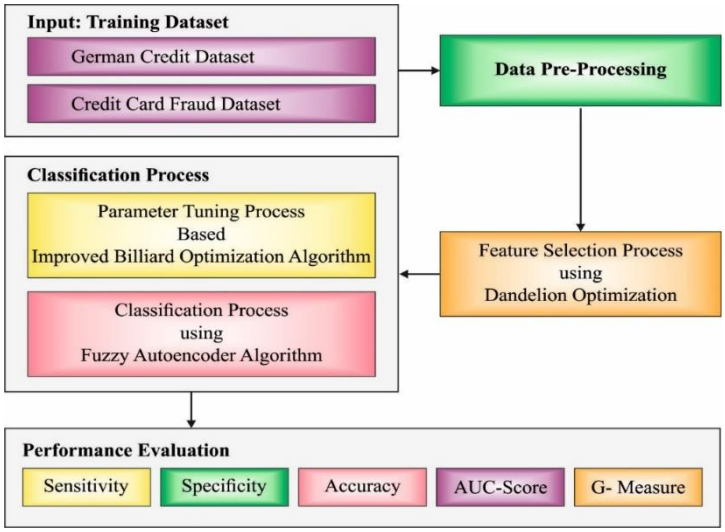

Figure 1. Overall working process of the ECFAE-CCFD technique.

3.1. Feature selection using DO-FS technique

The DO-FS method was executed to select the optimum features. The DO algorithm is based on the performance of determining the optimum reproduction place while dandelion seeds mature [22]. Also, it is highlighted that the flight behaviors of dandelion seeds are crucial in biological evolution. The longer-distance fight comprises three stages: descending, rising, and landing. The mathematical method of DO is discussed as follows.

The single objective DO technique with D parameters is formulated below:

$$\min f(X) \text{ ,} s.t. LB < X < UB, \tag{1}$$

where $f(X)$ signifies the objective function and $LB,\ UB \in R^D$ indicates the lower and upper boundaries of the parameter $X \in R^D$. Like other optimization techniques, DO comprises the following phases to resolve optimization problems.

DO randomly produces a candidate solution using Eq (2), where pop and Dim are correspondingly set to the population size and dimension parameter. $LB = (lb_1, lb_2, \cdots lb_{Dim})$ and $= (ub_1, ub_2, \cdots ub_{Dim})$ are the upper and lower limitations of the seed location.

$$X_{ij} = rand \times \left(ub_j - lb_j\right) + lb_j, i = 1,2, \ldots, pop, j = 1,2, \cdots, Dim, \tag{2}$$

$f(X_i)$ represents the fitness values of i^{th} seeds from the population and seed with smaller fitness is considered the best position for transmitting dandelion seeds, as X_{elite}:

$$f_{best} = \min\left(f(X_i)\right), X_{elite} = X\left(find\left(f_{best} = f\left(X_j\right)\right)\right), \tag{3}$$

where $find$ () represents an index with two equivalent values.

Dissimilar wind speeds and weather conditions define the increasing height of DO; hence, the weather was categorized into sunny and rainy.

Case1. During sunny days, the wind speed will be log-uniform distribution, meaning DO will have a higher probability to a distant area. Thus, dandelion seeds emphasize exploration on sunny days. This process can be mathematically modelled as follows:

$$X_i^{t+1} = X_i^t + \alpha^\star s_x^\star s_y^\star \ln Y^\star (X_s^t - X_i^t), \tag{4}$$

$$X_s^t = rand(1, Dim)^\star (UB - LB) + LB, \tag{5}$$

$$\alpha = rand()^\star \left(\frac{1}{T^2} t^2 - \frac{2}{T} t + 1\right), \tag{6}$$

$$r = \frac{1}{e^\theta}, s_x = r^\star \cos\theta, s_y = r^\star sin\theta, \tag{7}$$

where X_i^t represents the seed location at $the\ r$ iteration, the random location of the dandelion seed at the r iteration is represented as X_s^t, T denotes the maximal number of iterations, $\ln Y$ indicates a log-normal distribution followed by $\mu = 0,\ \sigma^2 = 1,\ \alpha$ shows the adaptive parameter, s_x and s_y indicate the dandelion seed lift module coefficient, and $\hat{I}$ denotes the arbitrary integer within $[-\pi, \pi]$.

Case2. During rainy days, DO cannot rise well with the wind, so DO emphasizes local neighborhood exploitation:

$$e = T^2 - 2T + 1, \tag{8}$$

$$\beta = 1 - rand()^{\star}\frac{1}{e}(t^2 - 2t - 1), \tag{9}$$

$$X_i^{t+1} = x_i^{t+1\,\star}\beta, \tag{10}$$

where β denotes the local adaptive parameter and T refers to the maximal number of iterations:

$$X_i^{t+1} = \begin{cases} X_i^t + \alpha^{\star}s_x^{\star}s_y ln\mathrm{Y}^{*}(X_s^t - X_i^t) randn < 1.5 \\ X_{\mathrm{i}}^{t\,*}\beta else \end{cases}, \tag{11}$$

where $randn$ shows the uniform distribution random integer.

Dandelion seeds emphasize global discovery in the decline phase. It facilitates the dandelion population and reflects the stability of decline to travel toward the preferred position for reproduction:

$$X_{mean\_t} = \frac{1}{pop}\sum_{i=1}^{pop} X_i, \tag{12}$$

$$X_i^{t+1} = X_i^t - \alpha^{\star}\beta_t^{\star}\left(X_{mean_t} - \alpha^{\star}\beta_t^{\star}X_i^t\right), \tag{13}$$

where X_{mean_t} denotes the mean place of the DO population in i^{th} iterations, and β_t refers to the Brownian movement.

Dandelion seeds focus on local neighborhood development in the landing stage. Based on the rising and descending phases, the DO arbitrarily chooses the landing site. The data around the existing elite seed can be utilized for local exploitation to approach the global optima:

$$X_i^{t+1} = X_{elite} + levy(\lambda)^{*}a^{*}X_{elite} - X_i^{t+1*}\delta)), \tag{14}$$

$$\delta = \frac{2t}{T}, \tag{15}$$

where X_{elite} stands for the better position of seeds at t iteration, T describes the maximal number of iterations, levy ($\hat{I}\rangle\rangle$) represents the function of Levy's fight, $\hat{I}\rangle\rangle = 1.5$, $\hat{I}$ linearly increases within [0,2]. In the presented DO-FS method, the fitness function (FF) is deployed to get a balance between the count of FSs from classifier accuracy (maximal) and every performance (minimal) achieved by using FS as follows:

$$Fitness = \alpha\gamma_R(D) + \beta\frac{|R|}{|C|}, \tag{16}$$

where $\gamma_R(D)$ denotes the classifier number of errors, α and β signify the two parameters equal to the effect of classifier quality and subset length, $\in$ [1,0] and $\beta = 1 - \alpha$, $|R|$ indicates the cardinality of the chosen subset, and $|C|$ indicates the overall count of features from the database.

3.2. Data classification using FAE model

The FAE model is used for the classification of CCF. Autoencoder (AE) was initially coined in the $1980s$ for dimensionality reduction with encoded and decoded parts [23]. During the encoded part,

the input layer $X = \{X_1, X_2, \cdots, X_N\} \in \mathbb{R}^{d\times N}$ is defined as a dimensionality decrease procedure as hidden state $H = \{H_1, H_2, \cdots, H_N\} \in \mathbb{R}^{m\times N}$ with weight linked matrix $W \in \mathbb{R}^{m\times d}$ and bias vector $B_1 \in \mathbb{R}^{m\times 1}$. During the decoder part, the HL reconstructs the input layer with $W \in \mathbb{R}^{d\times m}$ and $B_2 \in \mathbb{R}^{d\times 1}$ by minimalizing the loss function as follows:

$$\mathcal{L} = \frac{1}{d}||Y - X||^2. \tag{17}$$

The values of every node in the output layer were evaluated as follows:

$$Y = \sigma(W^T H + B_2), \tag{18}$$

where $\sigma(x)$ refers to a non-linear activation function, generally a logistic sigmoid function $\sigma(x) = 1/(1 + exp(-x))$. The HL values are attained using activation function $\sigma(x)$ and bias $\mathcal{B}_1$:

$$H = \sigma(WX + B_1). \tag{19}$$

Parameters $\Theta = \{W, B_1, B_2\}$ are used for minimizing the loss function. Particularly, the HL is considered an effective outcome of dimensionality reduction once the output reconstructs the input data. However, the conventional AE only follows the minimal reconstructed error in an undirected manner that is weaker to supervise. Therefore, classical AE is considered an unsupervised model. Figure 2 displays the structure of AE.

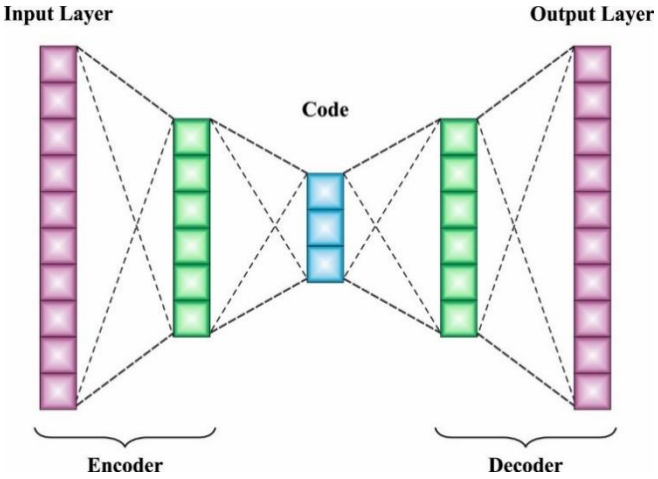

Figure 2. AE architecture.

The study focuses on extracting discriminatory representation by integrating fuzzy membership u_{ij} into the main function. Thus, the training method was guided by internally generated data, making the model self-supervised. FAE exploits AE's superior representation learning abilities and converts the information into another space with better discrimination by presenting a discriminative term with u_{ij},

$$\mathcal{L}(X, \Theta) = \min_{\Theta, C} \sum_{X_i \in X} [\frac{\eta}{2d}||X_i - Y_i||^2 + \frac{1-\eta}{2m}\sum_{j=1}^{K}\sum_{H_i \epsilon C_j} | |H_i - C_j||^2], \tag{20}$$

where $\Theta = \{W, B_1, B_2\}$ is the parameter of the model and η denotes a parameter adjusting for regulating the impact of the reconstructed loss and cluster-oriented loss. FAE should be trained in a self-supervised way and enhance the discriminatory learned features by presenting a clustering-oriented loss as to the presented method using fuzzy optimizer:

$$C_j = \frac{\sum_{H_i \epsilon C_j} u_{ij} H_i}{\sum_{H_i \epsilon C_j} u_{ij}}. \tag{21}$$

In the training process, the HL feature of the block was forced to cluster towards the block center, leading to the features with the best separability. Where the block center of HLs and in every iteration is denoted as C_j, and the discrimination of learned features can be improved by the similarity of instances from the block.

3.3. Hyperparameter selection using IBOA

The IBOA is exploited for the optimal selection of the hyperparameter values. Like other metaheuristics, the BOA technique has various disadvantages such as occasional instability and premature convergence [24]. Therefore, an improved version of BOA is introduced to resolve these drawbacks. To improve the efficacy in IBOA, a chaotic process can be employed by Lévy flight, which balances the exploration and exploitation:

$$Le(w) \cong w^{-(\xi+1)}, \tag{22}$$

$$w = A \times |B|^{-\frac{1}{\xi}}, \tag{23}$$

$$\sigma^2 = \left\{ \frac{\Gamma(1+\xi)}{\xi \Gamma((1+\xi)/2)} \frac{\sin(\pi\xi/2)}{2^{(1+\xi)/2}} \right\}^{\frac{2}{\xi}}, \tag{24}$$

where w describes the step size, Γ describes to the Gamma function, ξ represents the Lévy index within zero and two, $A, B \sim N(0, \sigma^2)$, and the value of ξ is assumed to be 3/2.

Therefore, the upgraded position of ordinary balls can be described as follows:

$$B_{n,s}^{new} = Le(\delta) \times (1 - PR)\left(B_{n,s}^{old} - P_{m,s}^{n}\right) + P_{m.s}^{n}, n = 1,2,3, \ldots N. \tag{25}$$

Whereas

$$A = a \times (2 \times r - 1), \tag{26}$$

$$B = C \times f(t) - B_{n,s}^{old}, \tag{27}$$

where $f(t)$ denotes the random location vector, and $a \in [0,2]$ and $r \in [0,1]$ represent the arbitrary variable. The pseudocode of IBOA is shown in Algorithm 1.

Algorithm 1: Pseudocode of IBO

Set N= number of balls';
M= number of variables;
K= number of pockets;
ET=escape threshold; iter $= 0$;
Initialization 2N balls and K pockets;
While (iter <iteration bound)
Evaluate the place of pockets and balls by employing the cost function;
Upgrade pocket memory and population;
Generate sets of ordinary and cue balls;
for every couple of balls
Select the target pocket by applying the roulette-wheel selection process;
End
Upgrade the position of the present normal balls;
Calculate the ordinary ball speed after collision;
Calculate the cue ball speed after collision;
Update the position of present cue balls;
If (rand $<$ ET)
Reconstruct the arbitrary size of balls;
End
Verify the boundary condition limitation and accurately define the ball range,
Iter = iter + 1;
Implement a chaotic Le'vy flight mechanism
End
Return to the optimal pocket for the outcome.

The fitness optimum becomes a main feature of the IBOA method. An encoding performance should be utilized to estimate the optimal of candidate effectiveness. The accuracy value is the main case utilized to design an FF.

$$Fitness = \max(P), \tag{28}$$

$$P = \frac{TP}{TP+FP}, \tag{29}$$

where TP is true positive values and FP is false positive values.

4. Results analysis

The proposed method is simulated utilizing Python 3.6.5 on PC i5-8600k, GeForce 1050Ti 4GB, 16GB RAM, 250GB SSD, and 1TB HDD. The parameter setting was specified as follows: batch size: 5, learning rate: 0.01, epoch count: 50, dropout: 0.5, and activation: ReLU.

In this section, the performance validation of ECFAE-CCFD method is tested under 2 databases a German credit database (http://archive.ics.uci.edu/dataset/144/statlog+german+credit+data) and a credit fraud detection database (https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud). The German credit dataset includes 1000 samples with a credit fraud detection dataset with 284807 samples as portrayed in Table 1.

Table 1. Description of two datasets.

Descriptions	German credit dataset	Credit fraud detection dataset
Source	UCI	Kaggle
# of instances	1000	284807
# of attributes	20	30
# selected attributes	13	19
# of class	2	2
Classes: Good/Bad	700/300	450/450

The classifier outcome of the ECFAE-CCFD method on the German credit database is exhibited in Figure 3. The confusion matrices obtained by the ECFAE-CCFD model on 70:30 of the TRPH /TSPH is illustrated in Figure 3(a) and (b). These findings specified that the ECFAE-CCFD algorithm can be appropriately detected and categorized with two classes. The PR outcome of the ECFAE-CCFD approach is depicted in Figure 3(c). The simulation value showed the ECFAE-CCFD algorithm has gained maximum PR solution on two classes. Moreover, the ROC examination of ECFAE-CCFD methodology is demonstrated in Figure 3(d). The outcomes showed that the ECFAE-CCFD method provides excellent performance with greater ROC outcomes on two classes.

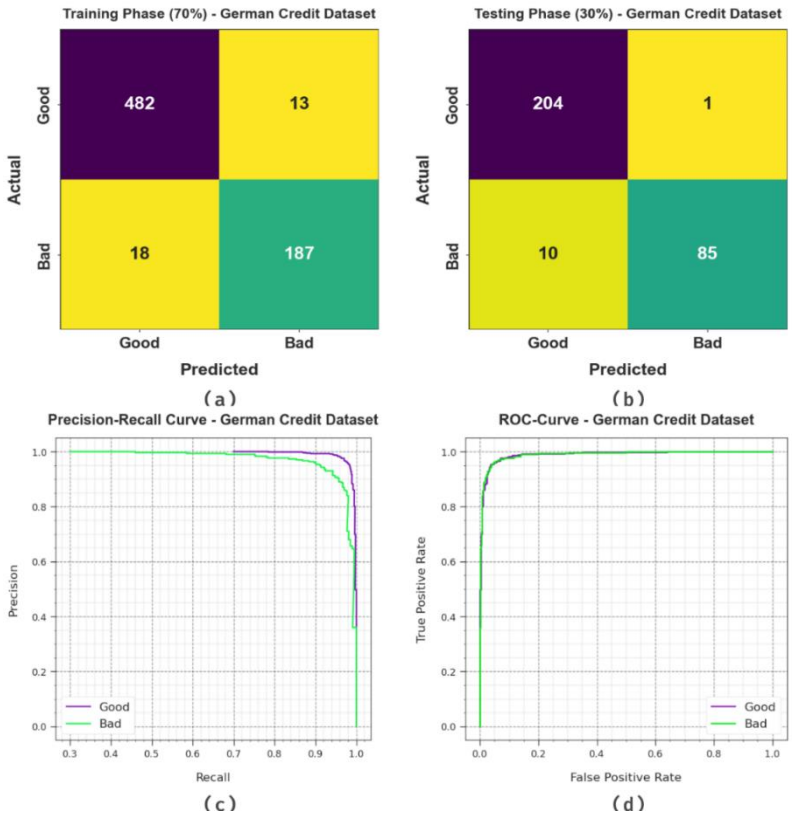

Figure 3. Performances on German credit dataset (a, b) confusion matrices, (c) PR_curve, and (d) ROC.

In Table 2, the stimulation value of the ECFAE-CCFD model under 70:30 of the German credit dataset. The simulation outcome of the ECFAE-CCFD technique states the good and bad samples.

According to 70% of TRPH, the ECFAE-CCFD approach achieves an average $accu_y$ of 94.30%, $sens_y$ of 94.30%, $spec_y$ of 94.30%, AUC_{score} of 94.30%, and $G_{measure}$ of 94.62%. Simultaneously, with 30% of TSPH, the ECFAE-CCFD method realizes an average $accu_y$ of 94.49%, $sens_y$ of 94.49%, $spec_y$ of 94.49%, AUC_{score} of 94.49%, and $G_{measure}$ of 95.72%.

Table 2. Classifier analysis of ECFAE-CCFD algorithm on German credit dataset.

Class	$Accu_y$	$Sens_y$	$Spec_y$	AUC_{Score}	$G_{Measure}$
TRPH (70%)					
Good	97.37	97.37	91.22	94.30	96.89
Bad	91.22	91.22	97.37	94.30	92.35
Average	**94.30**	**94.30**	**94.30**	**94.30**	**94.62**
TSPH (30%)					
Good	99.51	99.51	89.47	94.49	97.40
Bad	89.47	89.47	99.51	94.49	94.04
Average	**94.49**	**94.49**	**94.49**	**94.49**	**95.72**

The training accuracy TR_accu_y and VL_accu_y of the ECFAE-CCFD algorithm under the German credit dataset is described in Figure 4. The TL_accu_y can be calculated by evaluating the ECFAE-CCFD algorithm on the TR database while the VL_accu_y can be measured by calculating the effectiveness TS datasets. The experimental outcome exhibits that TR_accu_y and VL_accu_y upsurge with increasing epoch count. So, the efficiently of the ECFAE-CCFD approach is increased under datasets of the TR and TS with higher epoch count.

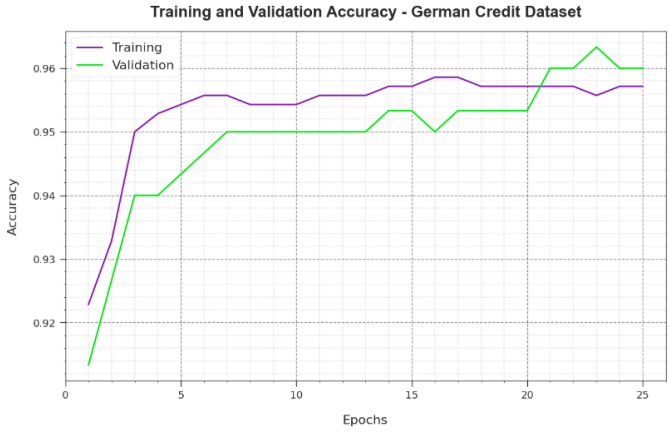

Figure 4. $Accu_y$ curve of ECFAE-CCFD algorithm on German credit database.

The TR_loss and VR_loss outcome of the ECFAE-CCFD method with German credit dataset are shown in Figure 5. The TR_loss describes the error among the original values and predictive outcomes on TR datasets. The VR_loss represents the performance metric of the ECFAE-CCFD algorithm on validation data. This experimental outcome indicates that the TR_loss and VR_loss decreased with the maximum epoch count. It showed the enriched outcomes of ECFAE-CCFD methodology and capability to create an exact classification. The minimized values of TR_loss and VR_loss demonstrate the enhanced outcomes of the ECFAE-CCFD model in capturing patterns and correlations.

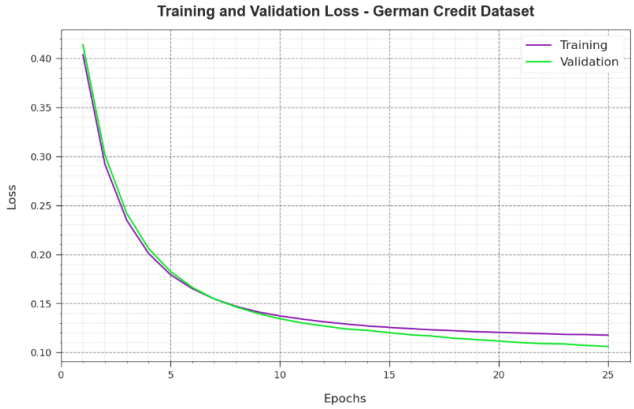

Figure 5. Loss curve of ECFAE-CCFD algorithm on German credit dataset.

The classifier outcome of the ECFAE-CCFD method on the credit fraud detection dataset is shown in Figure 6. The confusion matrices achieved by the ECFAE-CCFD system with 70:30 of the TRPH/TSPH is depicted in Figure 6(a) and (b). The accomplished findings outcomes showed that the ECFAE-CCFD technique can be precisely recognized and categorized the two classes. Next, the PR examination of the ECFAE-CCFD algorithm is shown in Figure 6(c). The simulation value showed that the ECFAE-CCFD algorithm had higher PR outcomes on two classes. Finally, the ROC curve of the ECFAE-CCFD methodology is represented in Figure 6(d). The ECFAE-CCFD model resulted in promising performance with enhanced ROC results on two classes.

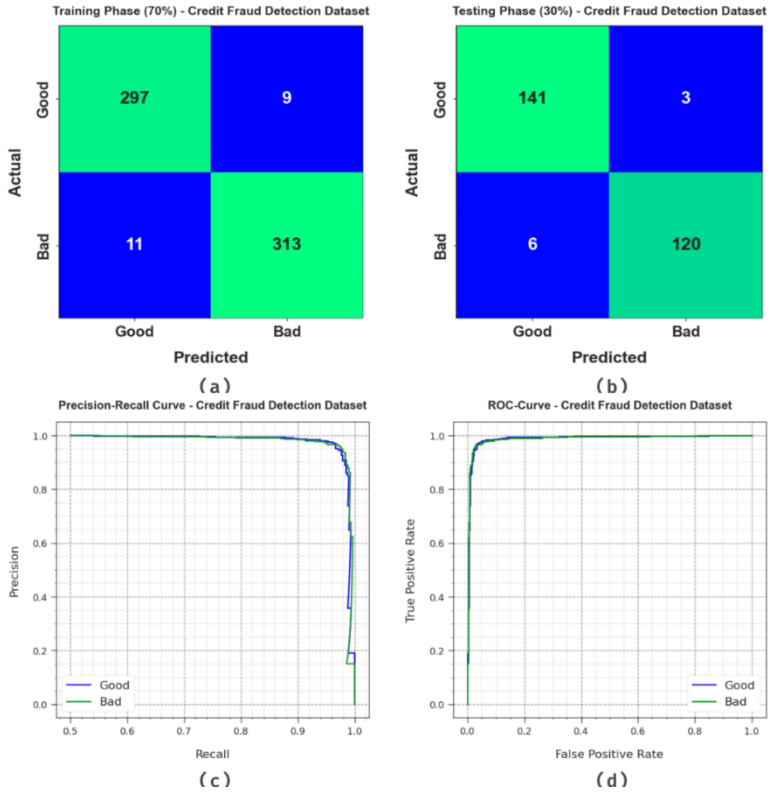

Figure 6. Performances on credit fraud detection dataset (a, b) confusion matrices, (c) PR_curve, and (d) ROC.

In Table 3, the experimental validation of the ECFAE-CCFD model under 70:30 of the credit fraud detection database. The simulation value outcomes of the ECFAE-CCFD approach state the good and bad samples. Based on 70% of TRPH, the ECFAE-CCFD system gets an average $accu_y$ of 96.83%, $sens_y$ of 96.83%, $spec_y$ of 96.83%, AUC_{score} of 96.83%, and $G_{measure}$ of 96.82%. Afterward, on 30% of TSPH, the ECFAE-CCFD technique accomplishes an average $accu_y$ of 96.58%, $sens_y$ of 96.58%, $spec_y$ of 96.58%, AUC_{score} of 96.58%, and $G_{measure}$ of 96.65%.

Table 3. Classifier outcome of ECFAE-CCFD algorithm on credit fraud detection dataset.

Class	$Accu_y$	$Sens_y$	$Spec_y$	AUC_{Score}	$G_{Measure}$
TRPH (70%)					
Good	97.06	97.06	96.60	96.83	96.74
Bad	96.60	96.60	97.06	96.83	96.90
Average	**96.83**	**96.83**	**96.83**	**96.83**	**96.82**
TSPH (30%)					
Good	97.92	97.92	95.24	96.58	96.91
Bad	95.24	95.24	97.92	96.58	96.39
Average	**96.58**	**96.58**	**96.58**	**96.58**	**96.65**

The training accuracy TR_accu_y and VL_accu_y of the ECFAE-CCFD technique on the credit fraud detection dataset is depicted in Figure 7. The TL_accu_y can be described by the calculation of the ECFAE-CCFD system with TR dataset while the VL_accu_y can be measured by calculating the outcomes on testing datasets. The experimental outcome shows that TR_accu_y and VL_accu_y are increased with increasing epoch count. Therefore, the effectiveness of the ECFAE-CCFD method can be increased on the datasets of TR and TS with maximum epoch count.

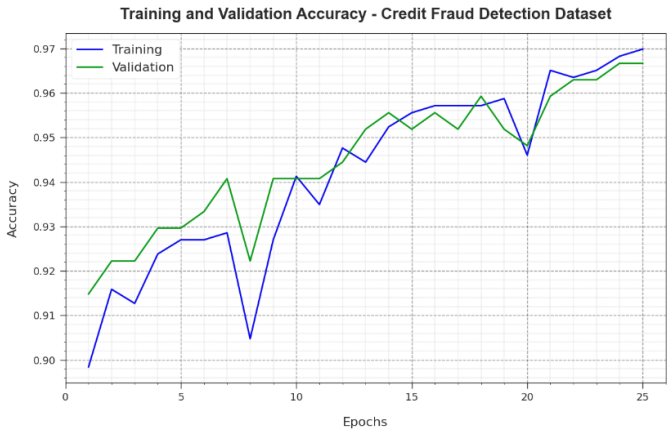

Figure 7. $Accu_y$ curve of ECFAE-CCFD algorithm on credit fraud detection dataset.

The TR_loss and VR_loss analysis of the ECFAE-CCFD algorithm on the credit fraud detection database can be seen in Figure 8. The TR_loss describes the error between the predictable outcome and original values on the dataset of TR. The VR_loss represents the effectiveness metric of the ECFAE-CCFD system with validation data. These experimental outcomes show that the TR_loss and VR_loss decreased with maximum epoch count. It demonstrated the enriched results of the ECFAE-

CCFD model and capabilities for producing correct classification. The minimum value of TR_loss and VR_loss shows the higher outcomes of the ECFAE-CCFD model in relationships and capturing patterns.

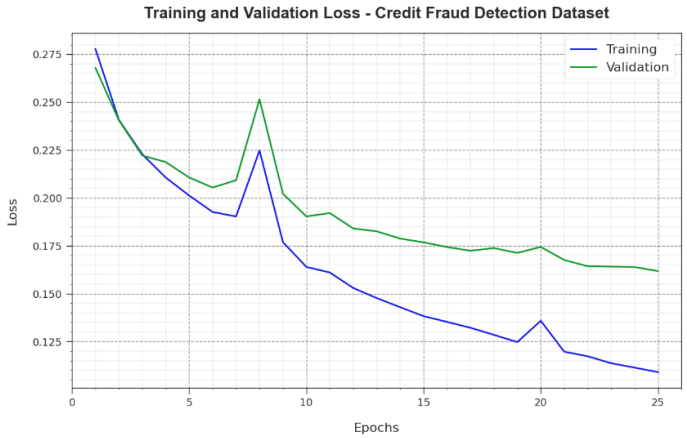

Figure 8. Loss curve of ECFAE-CCFD algorithm on credit fraud detection dataset.

In Table 4, a wide-ranging comparison analysis of the ECFAE-CCFD model is made with recent models [25]. Figure 9 investigates a brief outcomes analysis of the ECFAE-CCFD technique with respect to $accu_y$, AUC_{score}, and $G_{measure}$. Based on $accu_y$, the ECFAE-CCFD technique offers increasing $accu_y$ of 96.83% whereas the AdaBoost, LR, RF, SVM, ELM, IG-ELM, GAW, and ML-HFSICCFD techniques obtain decreasing $accu_y$ values of 80.23%, 80.24%, 83.23%, 90.70%, 80.71%, 79.29%, 89.80%, and 95.97% respectively. Also, with respect to AUC_{score}, the ECFAE-CCFD method offers an increasing AUC_{score} of 87%, whereas the AdaBoost, LR, RF, SVM, ELM, IG-ELM, GAW, and ML-HFSICCFD approaches achieve decreasing AUC_{score} values of 64%, 84%, 65%, 73%, 89%, 90%, 94%, and 96.83% respectively. Finally, in terms of $G_{measure}$, the ECFAE-CCFD approach achieves an increasing $G_{measure}$ of 88%, whereas the AdaBoost, LR, RF, SVM, ELM, IG-ELM, GAW, and ML-HFSICCFD systems gain lesser $G_{measure}$ values of 72.30%, 87%, 71.60%, 79.30%, 89.20%, 90.90%, 95.20%, and 96.82 % respectively.

Table 4. Comparison analysis of ECFAE-CCFD model with other algorithms [25].

Classifier	$Accu_y$	$Sens_y$	$Spec_y$	AUC_{Score}	$G_{Measure}$
AdaBoost	80.23	87.00	89.00	87.00	88.00
LR Model	80.24	62.50	83.70	64.00	72.30
RF Model	83.23	82.90	91.40	84.00	87.00
SVM Model	90.70	62.60	81.90	65.00	71.60
ELM Algorithm	80.71	71.00	88.50	73.00	79.30
IG-ELM	79.29	87.40	91.10	89.00	89.20
GAW Model	89.80	89.90	92.00	90.00	90.90
ML-HFSICCFD	95.97	94.50	96.10	94.00	95.20
ECFAE-CCFD	96.83	96.83	96.83	96.83	96.82

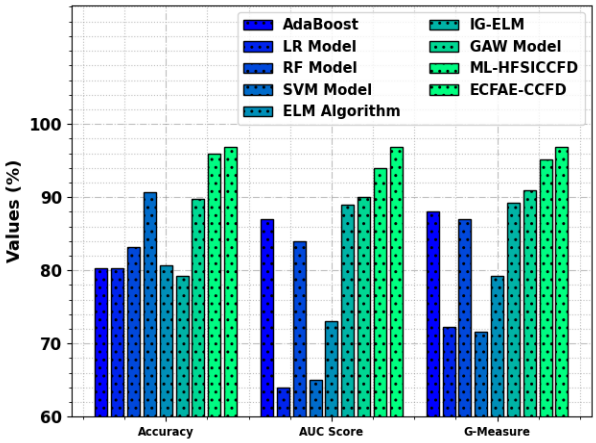

Figure 9. $Accu_y$, AUC_{score}, and $G_{measure}$ analysis of ECFAE-CCFD algorithm with other methods.

Figure 10 examines a brief results investigation of the ECFAE-CCFD method in terms of $sens_y$ and $spec_y$. Based on $sens_y$, the ECFAE-CCFD system attains enhanced $sens_y$ of 87%, whereas the AdaBoost, LR, RF, SVM, ELM, IG-ELM, GAW, and ML-HFSICCFD algorithms attain reduce $sens_y$ values of 62.50%, 82.90%, 62.60%, 71%, 87.40%, 89.90%, 94.50%, and 96.83% respectively. In addition, with respect to $spec_y$, the ECFAE-CCFD system obtains higher $spec_y$ of 89%, whereas the AdaBoost, LR, RF, SVM, ELM, IG-ELM, GAW, and ML-HFSICCFD algorithms gain minimal $spec_y$ values of 83.70%, 91.40%, 81.90%, 88.50%, 91.10%, 92%, 96.10%, and 96.83% respectively.

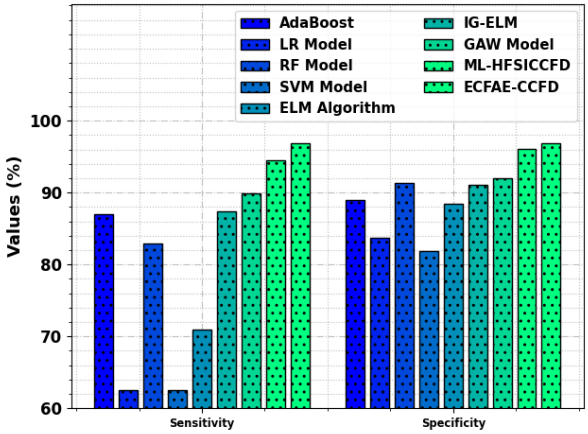

Figure 10. $Sens_y$ and $spec_y$ analysis of ECFAE-CCFD algorithm with other methods.

These outcomes highlighted the maximum efficacy of the ECFAE-CCFD method with other systems.

5. Conclusions

In this manuscript, we have presented the ECFAE-CCFD method. The major purpose of the ECFAE-CCFD model is to detect the presence of CCF in real time. To accomplish this, the ECFAE-

CCFD technique comprises data normalization, IBOA-based parameter tuning, FAE classification, and DO-FS-based feature subset selection. The ECFAE-CCFD method exploits the DO-FS technique for effectual selection of the features. Meanwhile, the FAE approach can be exploited for the recognition and classification of CCF. At last, the IBOA is applied for the optimum selection of parameters based on the FAE algorithm, increasing the classification accuracy. The simulation outcomes of the ECFAE-CCFD method could be examined on a benchmark open-access database. The obtained values display the promising performance of the ECFAE-CCFD system in terms of various measures.

The study could leverage a more comprehensive review of the practical applicability of the ECFAE-CCFD technique in practical scenarios. Specifically, insights into the adaptability of the method to diverse financial ecosystems, different scales of credit card transaction datasets, and the computational resources needed for real-time implementation could improve its relevance to real-time deployment. Furthermore, considering challenges such as data privacy regulations and incorporation with existing financial systems would provide a more detailed understanding of the feasibility and potential hurdles of the method in an actual operational context.

While the proposed method illustrates considerable developments in the field of automated fraud detection, it is crucial to consider its wider impact, especially in terms of ethical considerations. Automated fraud detection systems, such as ECFAE-CCFD, increase concern regarding bias, privacy, and transparency. The ethical implication might emerge from the wide usage of personal financial information and the potential for false positives impacting individuals. Transparency in the algorithm's decision-making process is vital to building trust, and this study could be beneficial for discussing how ECFAE-CCFD contributes or addresses to this ethical consideration. Furthermore, attention should be given to potential bias in the training data that might inadvertently perpetuate discriminatory outcomes. Since an automated fraud detection system plays a major role in a financial transaction, an ethical discussion surrounding the deployment of ECFAE-CCFD must emphasize the need for fair and responsible practices, ensuring that the benefits of improved fraud detection are balanced with ethical considerations to protect user trust and privacy.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors are thankful to the Deanship of Scientific Research at Najran University for funding this work under the General Research Funding program grant code (NU/DRP/SERC/12/16).

Conflict of interest

The authors declare that they have no conflict of interest. The manuscript was written through the contributions of all authors. All authors have given approval to the final version of the manuscript.

References

1. P. Roy, P. Rao, J. Gajre, K. Katake, A. Jagtap, Y. Gajmal, Comprehensive analysis for fraud detection of credit cards through machine learning. In: *2021 International conference on emerging smart computing and informatics (ESCI)*, 2021. https://doi.org/10.1109/ESCI50559.2021.9397029
2. J. Liu, X. Gu, C. Shang, Quantitative detection of financial fraud based on deep learning with combination of E-commerce big data, *Complexity*, **2020** (2020), 6685888. https://doi.org/10.1155/2020/6685888
3. E. Kim, J. Lee, H. Shin, H. Yang, S. Cho, S. K. Nam, et al., Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and deep learning, *Expert Syst. Appl.*, **128** (2019), 214–224. https://doi.org/10.1016/j.eswa.2019.03.042
4. A. RB, S. K. KR, Credit card fraud detection using artificial neural network, *Global Transit. Proc.*, **2** (2021), 35–41. https://doi.org/10.1016/j.gltp.2021.01.006
5. F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan, M. Ahmed, Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms, *IEEE Access*, **10** (2022), 39700–39715. https://doi.org/10.1109/ACCESS.2022.3166891
6. O. Voican, Credit card fraud detection using deep learning techniques, *Informatica Economica*, **25** (2021), 70–85. https://doi.org/10.24818/issn14531305/25.1.2021.06
7. X. Zhang, Y. Han, W. Xu, Q. Wang, HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture, *Inform. Sci.*, **557** (2021), 302–316. https://doi.org/10.1016/j.ins.2019.05.023
8. J. I. Z. Chen, K. L. Lai, Deep convolution neural network model for credit-card fraud detection and alert, *J. Artif. Intell. Capsule Netw.*, **3** (2021), 101–112. https://doi.org/10.36548/jaicn.2021.2.003
9. G. Pratuzaitė, N. Maknickienė, Investigation of credit cards fraud detection by using deep learning and classification algorithms. In: *11th International scientific conference "business and management 2020"*, 2020, 389–396. https://doi.org/10.3846/bm.2020.558
10. J. Forough, S. Momtazi, Ensemble of deep sequential models for credit card fraud detection, *Appl. Soft Comput.*, **99** (2021), 106883. https://doi.org/10.1016/j.asoc.2020.106883
11. A. R. Khalid, N. Owoh, O. Uthmani, M. Ashawa, J. Osamor, J. Adejoh, Enhancing credit card fraud detection: An ensemble machine learning approach, *Big Data Cogn Comput*, **8** (2024), 6. https://doi.org/10.3390/bdcc8010006
12. P. Raghavan, N. El Gayar, Fraud detection using machine learning and deep learning. In: *2019 international conference on computational intelligence and knowledge economy (ICCIKE)*, 2019, 334–339. https://doi.org/10.1109/ICCIKE47802.2019.9004231
13. T. K. Dang, T. C. Tran, L. M. Tuan, M. V. Tiep, Machine learning based on resampling approaches and deep reinforcement learning for credit card fraud detection systems, *Appl. Sci.*, **11** (2021), 10004. https://doi.org/10.3390/app112110004
14. A. Alharbi, M. Alshammari, O. D. Okon, A. Alabrah, H. T. Rauf, H. Alyami, et al., A novel text2IMG mechanism of credit card fraud detection: A deep learning approach, *Electronics*, **11** (2022), 756. https://doi.org/10.3390/electronics11050756

15. S. Sanober, I. Alam, S. Pande, F. Arslan, K. P. Rane, B. K. Singh, et al., An enhanced secure deep learning algorithm for fraud detection in wireless communication, *Wirel. Commun. Mob. Com.*, **2021** (2021), 6079582. https://doi.org/10.1155/2021/6079582
16. N. Nguyen, T. Duong, T. Chau, V. H. Nguyen, T. Trinh, D. Tran, et al., A proposed model for card fraud detection based on CatBoost and deep neural network, *IEEE Access*, **10** (2022), 96852–96861. https://doi.org/10.1109/ACCESS.2022.3205416
17. D. Almhaithawi, A. Jafar, M. Aljnidi, Example-dependent cost-sensitive credit cards fraud detection using SMOTE and Bayes minimum risk, *SN Appl. Sci.*, **2** (2020), 1574. https://doi.org/10.1007/s42452-020-03375-w
18. A. A. Taha, S. J. Malebary, An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine, *IEEE Access*, **8** (2020), 25579–25587. https://doi.org/10.1109/ACCESS.2020.2971354
19. V. R. Ganji, A. Chaparala, R. Sajja, Shuffled shepherd political optimization-based deep learning method for credit card fraud detection, *Concurr. Comp. Pract. E.*, **35** (2023), e7666. https://doi.org/10.1002/cpe.7666
20. A. Dhyani, A. Bansal, A. Jain, S. Seniaray, Credit card fraud detection using machine learning and incremental learning, In: *Proceedings of international conference on recent trends in computing*, Singapore: Springer, 2023, 337–349. https://doi.org/10.1007/978-981-19-8825-7_29
21. V. S. S. Karthik, A. Mishra, U. S. Reddy, Credit card fraud detection by modelling behaviour pattern using hybrid ensemble model, *Arab. J. Sci. Eng.*, **47** (2022), 1987–1997. https://doi.org/10.1007/s13369-021-06147-9
22. G. Hu, Y. Zheng, L. Abualigah, A. G. Hussien, DETDO: An adaptive hybrid dandelion optimizer for engineering optimization, *Adv. Eng. Inform.*, **57** (2023), 102004. https://doi.org/10.1016/j.aei.2023.102004
23. W. Yang, H. Wang, Y. Zhang, Z. Liu, T. Li, Self-supervised discriminative representation learning by fuzzy autoencoder, *ACM T. Intel. Syst. Technol.*, **14** (2022), 1–18. https://doi.org/10.1145/3555777
24. H. Ghafourian, S. S. Ershadi, D. K. Voronkova, S. Omidvari, L. Badrizadeh, M. L. Nehdi, Minimizing single-family homes' carbon dioxide emissions and life cycle costs: An improved billiard-based optimization algorithm approach, *Buildings*, **13** (2023), 1815. https://doi.org/10.3390/buildings13071815
25. I. D. Mienye, Y. Sun, A machine learning method with hybrid feature selection for improved credit card fraud detection, *Appl. Sci.*, **13** (2023), 7254. https://doi.org/10.3390/app13127254