

Computers, Materials & Continua DOI:10.32604/cmc.2022.021652 Article

An Improved Evolutionary Algorithm for Data Mining and Knowledge Discovery

Mesfer Al Duhayyim¹, Radwa Marzouk^{2,3}, Fahd N. Al-Wesabi⁴, Maram Alrajhi⁵, Manar Ahmed Hamza^{6,*} and Abu Sarwar Zamani⁶

¹Department of Natural and Applied Sciences, College of Community - Aflaj, Prince Sattam bin Abdulaziz University, Saudi Arabia

²Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Saudi Arabia

³Department of Mathematics, Faculty of Science, Cairo University, Giza 12613, Egypt

⁴Department of Computer Science, Faculty of Science & Art at Mahayil, King Khalid University, Saudi Arabia

⁵Department of Information Systems, Faculty of Science & Art at Mahayil, King Khalid University, Saudi Arabia

⁶Department of Computer and Self Development, Preparatory Year Deanship, Prince Sattam bin Abdulaziz University, Alkharj, Saudi Arabia

*Corresponding Author: Manar Ahmed Hamza. Email: ma.hamza@psau.edu.sa

Received: 08 July 2021; Accepted: 09 August 2021

Abstract: Recent advancements in computer technologies for data processing, collection, and storage have offered several chances to improve the abilities in production, services, communication, and researches. Data mining (DM) is an interdisciplinary field commonly used to extract useful patterns from the data. At the same time, educational data mining (EDM) is a kind of DM concept, which finds use in educational sector. Recently, artificial intelligence (AI) techniques can be used for mining a large amount of data. At the same time, in DM, the feature selection process becomes necessary to generate subset of features and can be solved by the use of metaheuristic optimization algorithms. With this motivation, this paper presents an improved evolutionary algorithm based feature subsets election with neuro-fuzzy classification (IEAFSS-NFC) for data mining in the education sector. The presented IEAFSS-NFC model involves data pre-processing, feature selection, and classification. Besides, the Chaotic Whale Optimization Algorithm (CWOA) is used for the selection of the highly related feature subsets to accomplish improved classification results. Then, Neuro-Fuzzy Classification (NFC) technique is employed for the classification of education data. The IEAFSS-NFC model is tested against a benchmark Student Performance DataSet from the UCI repository. The simulation outcome has shown that the IEAFSS-NFC model is superior to other methods.

Keywords: Educational data mining; feature selection; whale optimization; classification



This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1 Introduction

Educational Data Mining (EDM) is an emergent discipline, concern with emerging method to explore the different kinds of data which is derived from an educational setting, also utilize this method for better understanding student, and the setting where they learn [1]. The EDM is a multidisciplinary field of study which examine data mining, artificial intelligence and statistical modeling with the data made from an academic institutions [2]. EDM utilize computation way to handle explicate academic data considering the ultimate aim of examining academic enquiries. In order to make a country stands unique amongst other countries across the globe, the educational systems should experience a primary evolution by re-designing its structure. The hidden data and pattern from different data sources could be extracted by adapting the methods of data mining. For summarizing the performances of student with their credential, they explore the exploitation of data mining in the academic fields. All education institutions generate a huge amount of data yearly. Raw data could be significantly transferred through data mining model. The data attained from an education institution undergo scrutiny of different data mining methods [3]. The method detect the environments in which a students could be motivated better for leading a meaningful life so that they could contribute to society [4]. Weka, an effective data mining method is used for generating the result. The dramatic increase of education data [5] from heterogeneous source result an urgency for the EDM research. This might assist to encounter the goals and define certain objectives of education.

In machine learning as the dimensionality of the data increases, the amounts of data needed to provide a reliable analyses increase dramatically. Feature subset selections work by eliminating the feature which are redundant/aren't relevant. The subset of feature elected must follow the Occam Razor principle as well provide the significant result based on the objective function. In various cases this is an NP-hard (nondeterministic polynomial time hard) problems [6]. The size of the data to be treated have increased in last five years and thus feature selection is becoming a requirements before any kind of classification takes place. Different from feature extraction method, feature selection technique don't change the original representation of the data. One objective for both feature extraction and feature subset selection approaches is to evade over-fitting the data for making further analyses possible. The simplest is feature selection, where the amount of gene probes in a research is decreased by selecting only the most significant based on the conditions like higher level of activities.

Kou et al. [7] presented a 2-stage multiobjective Feature Selection (FS) technique which optimize the model classifier efficiency and amounts of feature. The outcomes demonstrated that the present techniques gained same classifier efficacy but highly decreasing the cardinality of the feature subsets. In Li et al. [8], a fast hybrid dimensional reduction technique to classifier is presented. A Multistrategy based FS was utilized for filtering out unrelated features. Also, the Grouped feature extraction was utilized for removing redundancy amongst features. The presented technique depicts outstanding performance and competitive classifier efficiency. Verma et al. [9] introduce a novel technique that implements 6 various DM classifier approaches becomes an ensemble manner utilizing gradient boosting, bagging, and AdaBoost classifications approaches for predicting the several class of skin diseases. Also, the feature essential technique was utilized for selecting vital 15 features that role an important play in forecast. The subsets of original datasets are gained behind electing only fifteen features for comparing the outcomes of utilized 6 ML approaches and ensemble manner as on the complete dataset.

In El-Hasnony et al. [10], a novel binary variants of the wrapper FSGWO and PSO is presented. The K-nearest neighbor (KNN) model with Euclidean separation matrices is utilized for finding an optimum solution. A tent chaotic map uses in avoided the technique in locked to the local optimum problems. In Arora et al. [11], a new Bolasso enabled Random forest technique (BS-RF) was presented for classifying borrowers as legitimate/defaulter. The stability of Bolasso is related to another FSs. BenSaid et al. [12] presented an online FS model that solves this issue. The presented OFS manner is named MOANOFS discover the present progress of online ML approaches and a conflict resolve manner (Automatic Negotiations) to the resolve of improving the classifier efficiency of ultra higher dimension database.

Quadri et al. [13] shows a C4.5 DT techniques for deploying a batch of features in hierarchical manner, also the models are employed many researchers since it is easier to classify the rule sets. Some of the common DT approaches such as C4.5 CART, J48 and so on. Murugananthan et al. [14] paced a new system for deriving an association rule that is employed to enhanced learning teacher and students' utilize K-means clustering model. An ANN is stated as an optimal approach i.e., used in mining educational detail. It is consider as an intelligent method which technique process according to the neurons that relates to each other and carry out in a combined way for generating a concluding result. Some other EDM models are also available in the literature [15–18].

The contributions of the study is given in the following. This paper presents improved evolutionary algorithm based feature subset selection with neuro-fuzzy classification (IEAFSS-NFC) for DM in the education sector. The presented IEAFSS-NFC model involves data pre-processing, FS, and classification. Besides, the Chaotic Whale Optimization Algorithms (CWOA) are employed to select the significant subsets feature using minimum cardinality to obtain maximum predictive efficiency. Then, Neuro-Fuzzy Classification (NFC) manner was utilized for classifying the educational data. The presented method was validated utilizing a benchmark Student Performance Dataset. Experimental result shows that optimum classifications of projected method on related techniques in various features.

The following section is organized below. The presented method is deliberated in Section 2. Then, it is documented in Section 3, as well the conclusion is derived in Section 4.

2 The Proposed IEAFSS-NFC Model

This method includes a group of 2 key steps i.e., classification and FS as displayed in Fig. 1. In the initial phase, transformation and format conversion arise as a pre-processing phase. Next, CWOA is performed for selecting a featured subsets from the academic information. Then, ANFC methods are used for classifying the features deceased data for identifying the class labels in a proper manner.

2.1 CWOA-FS Model

This study is applied with WOA which is mainly employed for FS. Here, WOA comes under the stochastic population-relied model which is stimulated from the bubble-net feeding in foraging behavior of humpback whale. The pseudocode representation of CWOA-FS is presented in Algorithm I. The case of bubble-net feeding is applied in a 6-shaped path. The WOA performs on the basis of 2 basic stages like exploration as well as exploitation. The initial phase named exploration works as whale's random searching task for prey.

All feature subsets are assumed as whale positions in WOA. Each subset is composed of the N features. By using the lower amount of features in all solutions, classification accuracy might be optimized. Additionally, all solution is verified on the basis of introducing fitness function (FF) which depends upon 2 major intentions: accuracy of a solution is reached by applying DNN classification process as well as the selected features of solution. Here, the model is initiated from the set of randomly

formed solutions named as a population. The application of proposed FF monitors each solution. From the population, the optimal solution implied as X* (prey) and major function in WOA is followed several times. For all iterations, the solutions are upgraded with their location on the basis of bubble-net attacking as well as searching prey models. In order to initialize the bubble-net attack, a minimum chance is concerned to choose one from the encircling technique (Eq. (2)) and spiral method (Eq. (6)) to update the location of whale's location.



Figure 1: Architecture diagram for the proposed model

	The task of	f enveloping the prey is denoted in Eqs. (1) and (2) .	
$\vec{\mathbf{D}}$	\vec{O} \vec{V}	T Y (A)	(1)

$$D = |\mathbf{C}.\mathbf{X}|(l) - \mathbf{X}(l)| \tag{1}$$

$$\vec{X}(t+1) = \vec{X}^*(t) - \vec{A}.\vec{D}$$
(2)

Whereas t is the current process, X^* implies the good solution reached, X denotes present solution, || indicates the proper measure whereas signifies an iteration by iteration multiplication. Also, the coefficient vectors are A and C that have been computed as follows.

$$\dot{A} = 2\vec{a}.\vec{r} - \vec{r} \tag{3}$$

$$\vec{C} = 2.\vec{r} \tag{4}$$

where *a* linearly minimizes from 2 to 0 and *r* represents the vector at arbitrary range of [0,1]. In order to improve the searching ability of WOA, even at the need of solutions to identify the randomly relied upon best solution location that has been explored, a randomly chosen solution is used on the basis of position updated. Hence, a vector A including random metrics more than 1 or minimum than -1

is used to create a solution to travel far away from best searching agent. It is expressed in Eq. (5) and (6).

$$\vec{D} = |\vec{C}.\vec{X_{rand}} - \vec{X}| \tag{5}$$

$$\vec{X}(t+1) = \overrightarrow{X_{rand}} - \vec{A}.\vec{D}$$
(6)

Whereas $\overrightarrow{X_{rand}}$ refers to an arbitrary whale chosen from current population. By the application of shrinking encircling pattern, an optimal trade-off among exploitation and exploration is essential. A random vector A which is present within the values of > 1 or < 1 has been applied. If A > 1, then the exploration process is carried out for exploring the adjacent randomly selected solution, whereas the nearby best solution undergoes exploitation while A < 1. The above process is repeated till the termination condition is convinced with more number of iterations.

Chaos is defined as semi randomness performance generated utilizing mathematically non-linear deterministic model which has 3 essential dynamic features as sensitivity, quasi stochastic, and ergodicity to new situation. The ability to change random parameters utilizing the value of chaotic map is defined by quasi stochastic. The capability of chaotic parameter for examining non-frequently whole situations in particular range was defined as ergodicity. The sensitivity to new situations is called as lesser variations in the premature state could lead to various performances. The efficacy of WOA is significantly improving with this feature. The chaotic map creates chaotic series that is employed to upgrade the whale place as well enhance the convergences speed and optimal solutions. While the efficacy of 10 chaotic maps. Therefore, for maximizing classifier accuracy and detects optimal feature subsets that minimize the count of FS, the logistic chaotic maps are fixed in optimization model in this work. In this work, the logistic chaotic maps were applied for alerting WOA variable.

2.2 Feature Extraction

In this work, solution is in the form of 1D vector with N elements, where N is the count features from raw dataset. Each field of as vector is constrained with the measure of "1" or "0". Rate "1" implies a similar feature is selected; otherwise values are fixed as "0".

The FF applied in this process is to obtain appropriate trade-offs between the selected features (lower) as well as classification precision (higher) reached by using predefined selected features. The FF is used to compute the solutions in the form of

$$Fitness = \propto \gamma_R(D) + \beta \frac{|R|}{|C|}$$
(7)

where $\gamma_R(D)$ is the classification error value of provided classier, |R| implies the cardinality of the selected subset while |C| denotes the entire features in dataset, α and β are 2 variables that refer the importance of subset length as well as classification quality. $\in [1, 0]$ and $\beta = (1 - \alpha)$ has been obtained.

Algorithm 1: Whale Optimization Based Feature Selection (CWOA-FS)				
Input: Number of Original Features (N)				
Output: Selected Features (S)				
1. Initialize the whale's population Xi $(i = 1, 2, 3,, n)$ as Features				
	(9			

(Continued)

Algorithm 1: Continued
2. Evaluate the fitness of all the whales
3. Fix X* as an optimal whale
4. while $(t < max_no. of_itera)$ do
4.1. For (each hunt whale) do
4.1.1. Upgrade a, A, C, l & p
4.1.2. If $(p < 0.5)$ then
4.1.2.1. If $(A < 1)$ then
4.1.2.2. The whale position is updating by moving towards best agent
4.1.2.3. Else
4.1.2.3.1. If $(A \ge 1)$ then
4.1.2.3.1.1. Choose the random whale X_{rand}
4.1.2.3.1.2. Then the whale position is updated
4.1.2.3.2. End If
4.1.2.4. End If
4.1.3. Else
4.1.3.1. If $(p \ge 0.5)$ then
4.1.3.1.1. Change the whale positions
4.1.3.2. End If
4.1.4. End If
4.2. End For
5. Note other search agents elsewhere in the searching area and revise it.
6. Compute the fitness of all search agents.
7. Upgrade X* when there is an optimum solutions
8. Mark the whales current position mapped as selected features (S)
9. $t = t + 1$

10. End while

2.3 ANFC Based Classification Model

The ANFC is assumed to be a hybrid technique that consists of maximum advantages as self adaptable as well as learning ability of NN and potentials of fuzzy method to consider the previous uncertainty as well as inaccuracy of real-time applications. The ability to deal with uncertainty and nonlinearity of data present in the given data helps to handle the classification process effectively and thereby leads to improved accuracy. The neuro-fuzzy modelling method has been compared with the extraction of mathematical information which describes the suitable character of this technique. By applying the ANFC method, a basic fuzzy model would be generated with the rules extracted from the input and output data [19]. Followed by, the NN is employed for tuning the basic method rules to produce excellent ANFC framework. Unlike ANN, it shows a higher capability in learning operation. Hence, it is used to modify the parameters present in membership function (MF) with automatic way as well as to reduce the error value in computing the fuzzy logic rules.

The ANFC structural method is a dynamic network that employs the supervised learning approach and is composed of the function of Takagi-Sugeno technology. Assume a batch of 2 two inputs x and y, and single outcome f of the structure. A collection of if-then rules are used as well as sample rules have been provided in the following.

• Rule 1: When x is A_1 and y is B_1 Then $f_1 = p_1 x + q_1 y + r_1$.

• Rule 2: When x is A_2 and y is B_2 Then $f_2 = p_2 x + q_2 y + r_2$.

where A_1, A_2 and B_1, B_2 are considered to be MF of all inputs x and y and p_1, q_1, r_1 and p_2, q_2, r_2 are linear variables that are within the final portion of Takagi-Sugeno method. ANFC has 5 layers as illustrated in Fig. 2. The primary as well as 4th layer contains an adaptive node, and residual layer is composed of static nodes.



Figure 2: Architecture of ANFC model

2.3.1 Layer 1

Each node adapts itself to a function parameter. The result obtained from all nodes is referred to as a degree of membership rates that is provided by input of MFs. For instance, MF used in this level is in the shape of bell as provided.

$$\mu_{Ai}(x) = \frac{1}{1 + \left|\frac{x - c}{a}\right|^{2b}}$$
(8)

where μ_{Ai} denotes the degree of MF for a fuzzy set Ai, whereas $\{a, b, c\}$ are assumed to be the attributes of MF which alter the MF structure.

2.3.2 Layer 2

All nodes present in the layer are predetermined or static and comprised with the multiplication operator Π . Each node of this layer describes the terminating power of all rules.

2.3.3 Layer 3

The single node within the layer is static or non-adaptable which is denoted by N. It generalizes the stopping potential as $\bar{w}_i = w_i / \Sigma w_i$.

2.3.4 Layer 4

All unique node of a layer is a dynamic node along with a function indicated as

$$\bar{w}_i f_i = p_i x + q_i y + r_i$$

(9)

The parameters of this layer can be described as consecutive variables.

2.3.5 Layer 5

All individual nodes in the layers are considered to be static or non-dynamic that helps to compute the entire result of each input signal from previous node as $\Sigma \bar{w}_i f_i$.

The initial and fourth layers are composed of tunable attributes in the training phase. The epoch value, MF as well as fuzzy rule value has to be selected accurately by using ANFIC technique, as it concludes in data over-fitting. This is achieved by the combination of least-squares and gradient descent (GD) models respectively.

3 Experimental Evaluation

In these sections, the accuracy measurement of the proposed IEAFSS-NFC models occurs. The presented method is executed by a MATLAB R2014a, run on a PC Intel RP Core [™] i7, 8GB RAM, 1TBHDD. The datasets employed to authentication and the equivalent result analyses is given as follows.

3.1 Dataset Details

The performances of the proposed IEAFSS-NFC models are authenticated by a benchmark datasets from UCI repository. The datasets hold an overall of 649 samples in the existence of thirty three attributes. Also, a overall of 10 class exists in the dataset. This detail is given in n Tab. 1.

 Table 1: Details of dataset

Value				
649				
33				
21				
[20]				

Tab. 2 illustrates the dispersal of class variable and stage and their equivalent range. The stage I lie in 16–20 range of stage 'Very Good'. Stage II, lie in 14–15 range of stage 'Good'. Stage III lie in 12–13 range of stage 'Satisfactory'. Stage IV lie in 10–11 range of stage 'Sufficient'. Stage V lie in of 0–9 range of stage 'Fail'.

Table 2: Class variable distribution (0–20)

Levels	Grade	Range
Ι	Very Good	16 to 20
II	Good	14 to15
III	Satisfactory	12 to 13
IV	Sufficient	10 to 11
V	Fail	0 to 9

3.2 Results Analysis

Fig. 3 depicts FS outcomes provided in the WOA models. Obviously, the figure show BC achieved from a group of twenty iterations. It is displayed that the CWOA-FS models have gained minimum BC of 0.00115. The values obviously showed an efficient FS result on the tested datasets.



Figure 3: Feature selection results in WOA model

Tab. 3 list the feature that is selected using CWOA-FS models on the employed datasets. It can be displayed that the CWOA-FS methods have selected a group of twenty features from an overall of thirty three attributes.

Methods	Selected features
CWOA-FS	1, 2, 4, 5, 7, 8, 10, 13, 14, 16, 18, 19, 20, 21, 22, 24, 26, 27, 31, 32

Table 3: Features chosen by CWOA-FS model

Tab. 4 and Fig. 4 provides the produced confusion matrix of the compared and proposed methods. Using the value presents in the confusion matrix, the equivalent classification result would be defined. Tab. 5 provide the result analyses given as the proposed method on the employed datasets. It is noted that the IEAFSS-NFC approaches have provided a recall of 98.07, maximal precision of 92.74%, F1 measures of 95.31%, precision of 92.29%, kappa values of 73.67%, and MCC of 74.61% correspondingly.

Fig. 5 showcases the ROC analysis of the IEAFSS-NFC technique with other existing techniques. From the figure, it is obvious that IEAFSS-NFC model has accomplished maximal ROC value of 98.0909.

Table 4: Comusion matrix					
Methods	ТР	TN	FP	FN	
IEAFSS-NFC	509	90	40	10	
NFC	450	86	99	14	
NN	381	42	168	58	
SVM	379	40	170	60	
DT	430	64	119	36	
RF	410	67	139	33	





Figure 4: Confusion matrix analysis of different models

Tab. 5 and Figs. 6-8 shows the comparison of the outcomes provided by the existing and proposed methods in numerous performances measure. The figure obviously state that the SVM models show minimum classification accuracy also obtain a recall of 92.27%, least accuracy of 69.03%, MCC of 33.96%, kappa value of 6.23%, and F1-measure of 76.72%. Then, the NN techniques have provided rather optimal result on SVM models and achieve an accuracy of 69.39%, F1-measure of 77.12%, recall of 86.78%, MCC of 8.80%, and kappa value of 7.86%. Afterward, RF classifiers have attempted for well performing and end-up by a least accuracy of 74.68%, F1-measure of 82.66%, recall of 92.55%, kappa value of 29.07% and MCC of 32.33%. At the same time, controllable result is given to the DT

models with the accuracy of 78.32%, F1 measure of 76.72%, recall of 92.27%, MCC of 33.96%, and kappa value of 31.59%.

Methods	Precision	Recall	Accuracy	F1-measure	MCC	Kappa
IEAFSS- NFC	92.71	98.07	92.29	95.31	74.61	73.67
NFC	81.96	96.98	82.60	88.84	54.36	50.43
NN	69.39	86.78	65.10	77.12	08.80	07.86
SVM	69.03	86.33	64.50	76.72	06.97	06.23
DT	78.32	92.27	76.10	84.72	33.96	31.59
RF	74.68	92.55	73.50	82.66	32.33	29.07

Table 5: Comparison study of IEAFSS-NFC technique under several measures



Figure 5: ROC analysis of IEAFSS-NFC technique

Reasonable performances are provided as a NFC model that offer a higher accuracy of 81.96%, F1-measure of 88.84%, recall of 96.98%, MCC of 54.36%, and kappa value of 50.43%. But, the presented IEAFSS-NFC models have exceed the related method with the attainment of maximal precision of 92.74%, F1 measures of 95.31%, recall of 98.07%, kappa values of 73.67% and MCC of 74.61%. This value ensures the dominance of the presented method on related techniques.



Figure 6: Precision and recall analysis of the IEAFSS-NFC technique



Figure 7: Accuracy and F1-measure analysis of the IEAFSS-NFC technique

On examining the results interms of accuracy, the table values noted that SVM models have showed minimum classification accuracy i.e., guaranteed with 64.50% of accuracy. T, the NN models have given rather high classifiers result on SVM through providing an accuracy of 65.10%. Afterward, the RF classifiers have displayed rather reasonable classifiers result by attaining 73.50% of accuracy. Simultaneously, DT classification models have given controllable classification by attaining 76.10% of accuracy. At the same time, NFC models have provided competitive classifiers outcome by providing 82.60% of accuracy. However, the proposed IEAFSS-NFC models have provided a maximal classifiers

92.29% of accuracy. Clearly, the abovementioned figures and tables showed the excellent features of the proposed method on the related model.



Figure 8: MCC and Kappa Analysis of the IEAFSS-NFC technique

4 Conclusion

The study has presented an efficient IEAFSS-NFC model for assessing the students' efficiency at the time of learning procedure. The proposed model includes a group of 2 main stages such as classification and FS models. In the earlier stage, CWOA is performed for selecting a feature subsets from the academic details. Then, ANFC models are used for classifying the feature reduction data to find a class labels in a proper manner. The result of IEAFSS-NFC models were authorized by a benchmark datasets from UCI repository. Stimulation outcome shows that IEAFSS-NFC techniques have provided a maximal accuracy of 92.74%, precision of 92.29%, recall of 98.07%, kappa values of 73.67%, MCC of 74.61% and F1 measure of 95.31% correspondingly. After this, the presented method could be realized in a realtime applications.

Funding Statement: The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work under grant number (RGP2/209/42) received by Fahd N. Al-Wesabi, www.kku.edu.sa. This research was funded by the Deanship of Scientific Research at Princess Nourah bint Abdulrahman University through the Fast-Track Path of Research Funding Program.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

- S. N. Mohanty, E. L. Lydia, M. Elhoseny, M. M. G. Al Otaibi and K. Shankar, "Deep learning with LSTM based distributed data mining model for energy efficient wireless sensor networks," *Physical Communication*, vol. 40, pp. 1–9, 2020.
- [2] S. K. Lakshmanaprabu, K. Shankar, D. Gupta, A. Khanna, J. J. P. C. Rodrigues *et al.*, "Ranking analysis for online customer reviews of products using opinion mining with clustering," *Complexity*, vol. 2018, pp. 1–9, 2018.
- [3] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1432–1462, 2014.
- [4] K. Shankar, M. Ilayaraja, P. Deepalakshmi, S. Ramkumar, K. S. Kumar *et al.*, "Opinion mining analysis of e-commerce sites using fuzzy clustering with whale optimization techniques," *Expert Systems in Finance*, 1st ed., Routledge, London: CRC Press, pp. 67–79, 2019.
- [5] S. K. Lakshmanaprabu, K. Shankar, M. Ilayaraja, A. W. Nasir, V. Vijayakumar *et al.*, "Random forest for big data classification in the internet of things using optimal features," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 10, pp. 2609–2618, 2019.
- [6] N. Krishnaraj, S. Krishamoorthy, S. V. Lakshmi, C. S. R. Priya, V. Dahiya et al., "Big data based medical data classification using oppositional gray wolf optimization with kernel ridge regression," *Applications of Big Data in Healthcare*, 1st ed., US: Academic Press, pp. 195–214, 2021.
- [7] G. Kou, Y. Xu, Y. Peng, F. Shen, Y. Chen *et al.*, "Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection," *Decision Support Systems*, vol. 140, pp. 113429, 2021.
- [8] M. Li, H. Wang, L. Yang, Y. Liang, Z. Shang *et al.*, "Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction," *Expert Systems with Applications*, vol. 150, pp. 1–29, 2020.
- [9] A. K. Verma, S. Pal and S. Kumar, "Prediction of skin disease using ensemble data mining techniques and feature selection method—A comparative study," *Applied Biochemistry and Biotechnology*, vol. 190, no. 2, pp. 341–359, 2020.
- [10] I. M. El-Hasnony, S. I. Barakat, M. Elhoseny and R. R. Mostafa, "Improved feature selection model for big data analytics," *IEEE Access*, vol. 8, pp. 66989–67004, 2020.
- [11] N. Arora and P. D. Kaur, "A bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment," *Applied Soft Computing*, vol. 86, pp. 1–27, 2020.
- [12] F. BenSaid and A. M. Alimi, "Online feature selection system for big data classification based on multiobjective automated negotiation," *Pattern Recognition*, vol. 110, pp. 107629, 2021.
- [13] A. M. Shahiri, W. Husain and N. A. Rashid, "A review on predicting student's performance using data mining techniques," *Procedia Computer Science*, vol. 72, pp. 414–422, 2015.
- [14] V. Murugananthan and B. L. ShivaKumar, "An adaptive educational data mining technique for mining educational data models in elearning systems," *Indian Journal of Science and Technology*, vol. 9, no. 3, pp. 1–5, 2016.
- [15] P. M. Arsad, N. Buniyamin and J. A. Manan, "A neural network students' performance prediction model (NNSPPM)," in 2013 IEEE Int. Conf. on Smart Instrumentation, Measurement and Applications (ICSIMA), Kuala Lumpur, Malaysia, pp. 1–5, 2013.
- [16] N. T. N. Hien and P. Haddawy, "A decision support system for evaluating international student applications," in 2007 37th Annual Frontiers in Education Conf. - Global Engineering: Knowledge Without Borders, Opportunities Without Passports, Milwaukee, WI, pp. F2A-1-F2A-6, 2007.

- [17] E. A. Amrieh, T. Hamtini and I. Aljarah, "Preprocessing and analyzing educational data set using X-aPI for improving student's performance," in 2015 IEEE Jordan Conf. on Applied Electrical Engineering and Computing Technologies (AEECT), Amman, Jordan, pp. 1–5, 2015.
- [18] E. A. Amrieh, T. Hamtini and I. Aljarah, "Mining educational data to predict student's academic performance using ensemble methods," *International Journal of Database Theory and Application*, vol. 9, no. 8, pp. 119–136, 2016.
- [19] P. Borkar and L. G. Malik, "Acoustic signal based traffic density state estimation using adaptive neurofuzzy classifier," in 2013 IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE), Hyderabad, India, pp. 1–8, 2013.
- [20] Dataset. 2014. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/student+performance.