An Enhanced Technique to Improve the Performance Classification in Data Mining Using Recurrent Distribution Correlated Optimization Data Mining From Social Network Using Neural Network

NAVEEN KUMAR ML

Dept. Of Electrical and Electronics Engineering, North East Frontier Technical University [NEFTU], Medog, Aalo, Arunachal Pradesh, India

Abstract- ***To enhance the classification performance and to reduce the time complexity for large amount of sensor data analysis, optimization method can be implemented to select best attribute among the overall feature database. This can be achieved by using the Recurrent Distribution Correlated Optimization (RDCO) algorithm to find the relevancy between the feature of query data and from the entire dataset and selects the best optimal features. In this, the data can be preprocessed and clustered by using the Maximum Possibility Combination (MPC) based clustering algorithm. To find the matching feature, the Multi-Block Convoluted Learning (MBCL). With this system, first the pre-processed feature is matched with the pattern by using MBCL to find the type of data without directly passed into the whole dataset. From that type identified result, the similarity between the matched result and overall dataset is retrieved by using the RDCO method to display all matched result from the bulk dataset with better classification result.***

Indexed Terms- ***Classification, Data mining, Data prediction, Feature extraction, Query optimization***

I. INTRODUCTION

Utilizing conventional data analysis techniques, the amount of data on the Internet is exponentially growing, and its influence on many facets of social production and daily life is becoming more and more obvious. On the basis of this, clustering optimization methods and data mining techniques are produced. [1] To execute the data mining process, we first establish the mining task and then choose the appropriate mining algorithm. Human-computer interaction is used often during the mining process. It mostly entails identifying issues, creating data mining libraries, analyzing data, gathering data, creating models, assessing models, and putting those models into practice. Professional expertise in the application field, database, data warehouse, or other information repositories is integral to the entire data mining process. [2]

In order to find patterns and learn more about such patterns, data mining often refers to mining or delving deeply into data that is in various forms. Large data sets are initially sorted in the data mining process before patterns are found and connections are made to conduct data analysis and address issues. [3]

- *Classification*

Finding a model that explains and differentiates data classes and ideas is the process that makes up the data analysis task. On the basis of a training set of data that includes observations and whose category membership is known, classification is the issue of determining which of a set of categories (subpopulations), a new observation belongs to [4]. The following steps make up a fundamental categorization model:

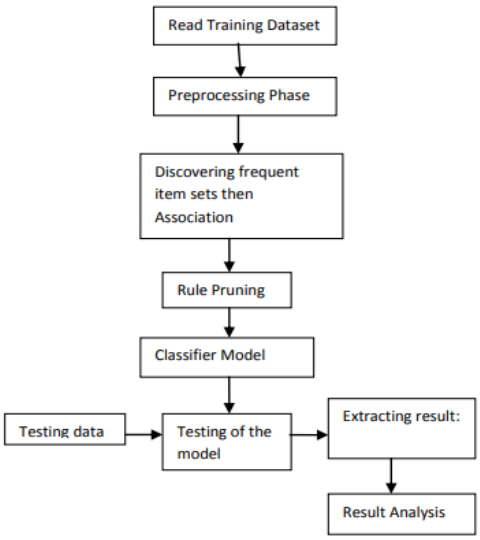

Fig. 1: Basic model classification process

- Pre-processing: By allowing the model to learn using the provided training data, several algorithms are employed to construct a classifier. For reliable outcome prediction, the model must be trained.
- Classification Step: The classification step involves building a model to predict class labels, testing it on test data, and estimating how accurate the classification rules are.
- Training and Testing: • Training and Testing: If a person is seated beneath a fan and it starts to fall on him, he should get away to avoid hurt. Thus, he is practising moving away at this point. When a person is being tested, the system is either favourably or negatively assessed depending on whether they move away when they notice a large item approaching their way or about to land on them. In order to get accurate and superior outcomes, the data must also be trained. [5]

In this proposed work we are using hierarchical type of clustering. We test our approach against a well-known data set containing a range of classes, instances, and attributes in order to determine how effective the suggested algorithm is. [6]

Objectives of this paper are as follows:

- To develop an enhanced model of data prediction database management system.
- To enhance the data prediction system by using the MPC clustering system.
- To analyse the performance of proposed prediction model compared with traditional model of data prediction techniques.

The paper is alienated into six main segments. Section I deals with introduction to research work with elementary explanation of concepts. Section II is about the related work and to evaluate the review for new proposed work. The section III deals with the explanation of proposed work mainly with the proposed methodology and algorithm. The section IV and V described about the experimentation and results. Last section VI contains conclusion of paper and future suggestions of the proposed work.

II. RELATED WORK

The choice of data mining technique will have an impact on the effectiveness and quality of the findings since different approaches have different functional features and suitable domains. Multiple technologies are typically used to create complimentary benefits throughout the actual application process.

- The majority of the computations, according to Jorge, are fundamentally incorrect. Jorge's discoveries gave rise to CN volatile (CNV), which had minimal accuracy loss compared to the most sophisticated accelerator [7], removed the majority of these incorrect operations, enhanced performance, and consumed less energy.
- Han put up the idea of an EIE (energy-efficient inference engine) [8]. Wang employs a neural network model in the microwave design process through a process called training to deliver immediate responses for the taught tasks after learning and extracting microwave data. The development of neural network models for microwave applications has two major challenges: selecting an appropriate neural network topology and training method [9].

- A time segmentation framework and particular course performance metrics gathered from course reports which are applied in estimating the students' function. The key objective of this work is to present methods for EDM and combine into a web-based system and predict the poor students [10].
- In [11] the case study analysis in educational data by examining the data with the help of the DM model. A researcher applies the classifier which concentrates on drop-out forecasting UG and diploma students. In order to detect the drop-out classification, academic details are required. The actual data of student's academic data enrolls in university from past decades.
- The major objective of [12] study is to deploy a classifier under the application of locally produced students' features for exact performance detection. Followed by, students' features are gathered from distinct sources that are preprocessed that are further established using WEKA for FS and consequently for learning and testing. The NB classification model has emerged as a précised classifier and executed in predictor tool.
- [13], the ensemble models are combined with the effective phenomenon in classification as well as prediction where the developers make use of boosting method and deploy the exact prediction pedagogical approach and the exposed nature of newly presented technique in EDM.
- [14] Predictive analysis was carried out to compute the cumulative grade point average (CGPA) for the final year of engineering students at Nigerian Universities using the application of the study plan, the admission date, and the GPA for the first three years of study, which are referred to as inputs for Konstanz Information Miner (KNIME) related DM method.

III. PROPOSED WORK

The identification process of anomaly detection in wireless network and grouping represents a dynamic update of data from the network database to analyse the state of anomaly prediction in wireless network and mining system for industrial dataset. This will be update for a period of time interval that processed in frequent order. [15] There are several methods in classification and analyse the database contents like neural techniques and other machine learning technics. In that, sequence pattern method with the several Neural Network were most commonly used to classify and match the relevant features from database. Since, for the huge amount of node characteristics in data, it struggles in predicting class with proper attributes of feature vector. These searching process predicts the relevant feature that compare the input query with the database in high-speed analysis of large sensor data. [16]

In the recent research work, the matching prediction is complicated due more irrelevant features present in the database. To improve the performance of feature identification in the searching process, machine learning technique helps to predict the best matching between the query data and feature sets in database. [17] There are several types of machine learning techniques like Support Vector Machine, Relevant Vector Machine and other methods. Since, in the recent days, the Neural Network and Deep Learning technique takes place a major role of data analysis and match prediction among the bulk amount of raw data. [18]

In this, the data can be preprocessed and clustered by using the MPC based clustering algorithm. To find the matching feature, the MBCL. With this system, first the pre-processed feature is matched with the pattern by using MBCL to find the type of data without directly passed into the whole dataset. [19]

A. Proposed Model

Proposed model considers these modules as preprocessing, grid formation, pattern formation, classification and data prediction. [20] Fig. 2 presented the stepwise formation of proposed model which mainly includes following steps.

1) Pre-processing: This step includes the selection of data input and process for its testing.
2) MPC segement: this steps incvolves the formation of grids using above part of clustering. This steps uses the input coming from the feature selection part which is perfromed using MPC.
3) Traing and testing of features: This step perfrom the testing and training of featured data using RDCO method. The input data used as the grid

formed in above step. The output of this step used as neurons of input of neural network.

4) Classification using MBCL: output of features database from the training of dataset and output from MBCL classifier prepares the classification results that shows the data for prediction.
5) Data prediction: final output of the whole process are in form of data prediction presented through this step.

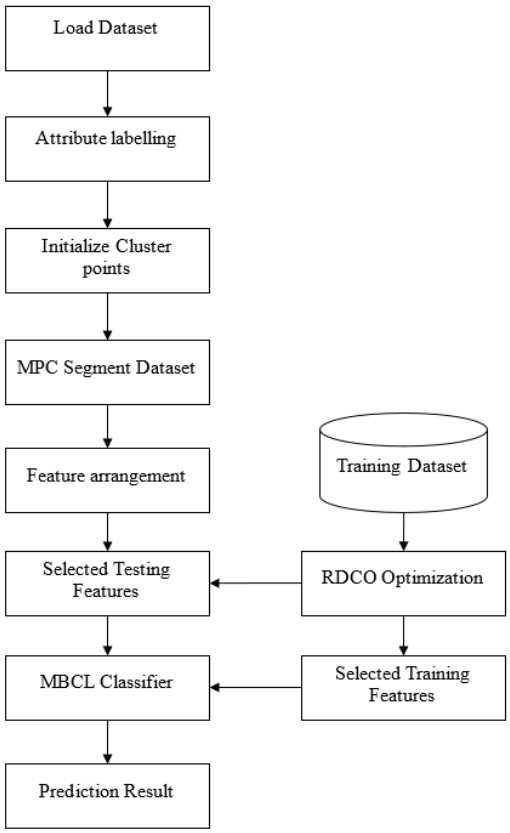

Fig. 2: Flow diagram of proposed model

B. Proposed techniques

There are two major techniques used for proposed work. [21] That are as follows:

1. Maximum Possibility Combination (MPC):

This technique explains the working process of clustering algorithm for feature analysis [22]. The steps of this technique are as follows:

Input: Input Data Matrix [(M_{NID})]

Compute the initial clustering of data:

Input: Data stream D_j

Output: Matching M ⊆ E

For Loop run (M = 0 to ∅)

Update P = {D_1, ..., D_k} Data-maximal set

L = length of the Data Stream

$$M = M \oplus (P(1) \cup P(2) \cup ... \cup P(k)) \quad (1)$$

Until P = ∅ Loop

Repeat loop 'M'

End For 'M'

2. Recurrent Distribution Correlated Optimization (RDCO):

Testing and training of datasets using technique [23];

Input: Data $\{N_i\}$, Cost of Particles c_{ij}

Output: Best selection of Parameters

Run the For Loop for 'm' number of data-streams in a database as of iteration 1 to m.

Initialization of the weight value as γ of data-stream which can be signified as

$$\gamma = \{N, L\} \quad (2)$$

where,

'N' = the set of data in a database (3)

'L' = length / distance between each data (4)

Run the For Loop for 'i' number of data-counts in a database as of iteration 1 to n.

Run the For Loop for 'j' number of data-counts in a database as of iteration 1 to l.

Where '*l*' is the number of streams

Produce data-stream flow $\{f_{i,j}^k\}$ for 'k' no. of trials

Update $\gamma_{i,j}^k = \gamma_{i,j}(f_{i,j}^k), \forall(i,j) \in L$

Compute route choice probabilities $\{y_{i,j}^k\}$ in the database architecture as

$$\{y_{i,j}^k\} = \sum_{s,d} \sum_t h_{s,d} \times P(r|C_n)\left(v_{s,d}(f_{i,j}^k)\right) \times a_{i,j}^r \quad (5)$$

where, (s, d) denotes the source-destination pair.

$v_{s,d}$ – Anomaly weight of network image

$a_{i,j}^{r}$ – denotes the no. of transmissions to data in (i, j) seems in the construction for the coverage-size of 'r'.

Update flow pattern,

$$f_{i,j}^{k+1} = f_{i,j}^{k} + \alpha_n \times \left(y_{i,j}^{k} - f_{i,j}^{k}\right) \quad (6)$$

Check convergence for respectively k+1 value

Calculate $G(y) = max(y_{i,j}^{k})$. (7)

Find highest possible point for identifying selection length

Calculate $L(j) = \frac{1}{n}\sum_{x=1}^{n}\left\|N\left(f_{i,j}^{k}\right) - G(y)\right\|$ (8)

Find the distance vector between each data.

If $\lambda < L$, then // λ defines the particle strength of data

Calculate

$$\Delta_{(j)} = \Delta_{j-1} + \mu \times \partial L/\partial W_i^l \quad (9)$$

Find optimum data with neighboring distance assets from the index of data.

Calculate $c_{i,j} = W_j^l + \Delta_{(j)}$ (10)

Update the cost value with Anomaly in table

Continue.

Else

Image acknowledgement and estimation training Anomaly parameters with data parameters

Continue loop.

End If

End For '*j*' and '*i*'

Select the best selection from the updated table '$y_{i,j}^{k}$'

Update database weight and architecture.

End For

3. Multi-block Convoluted learning (MBCL):

Using eq. (1) and (6), compute data stream

Input: Data stream D_j

Output: Identified defection E_j

Run For Loop 'j' for the data-size 'm'

j = 1 to m

Initialization of parameters 'K' and 'S' as an empty array of bits.

Initialization of the arbitrary integer value for the size of data stream (i = length(D)) and 'sk' and 'vk'.

The arbitrary hash parameter can be produced by the below eq. (11) of $H_a(j)$.

$$H_a(j) = \left|\left(\frac{aK \bmod W}{\left(\frac{W}{M}\right)}\right)\right| \quad (11)$$

Where, 'M' signify the max length of data-size.

IV. SIMULATION SCENARIO

Experimentation process have demonstrated the process of proposed work. [24] Following parameters are used to do the experimentation work:

Table 1: List of Parameters used in simulation

Parameters	Values
Mac protocol	802.15.4
Number of nodes	100
Cluster Head location	(150, 150) m
Time slot width	5 sec
Routing Protocol	AODV
Node's transmission range	60 m
Simulation time	06h:30m:50s
Total time slot	4690
Network size	300 m × 300 m
Transport Protocol	UDP
Event sensor reading (Temperature)	N (29.3112, 4.5588)
Percentage of measurement faulty nodes	50%, 40%, 30%, 20%, 10%, and 5%
Faulty sensor reading (Temperature)	N (29.3112, 4.5588)
Normal sensor reading (Temperature)	N (28.1273, 1.0952)

Faulty sensor reading (Humidity)	N (78.4943, 11.3831)
Normal sensor reading (Humidity)	N (59.6504, 9.7391)
Event sensor reading (Humidity)	N (78.4943, 11.3831)

V. RESULT

Comparative result analysis has confirmed the development of proposed work by using the mentioned dataset as shown in below tables and figures.

Table 2: Performance evaluation for different parameters

Metrics	PCC with SVM	Proposed
Accuracy	93.77%	97.25%
Precision	95.09%	98.57%
Sensitivity	94.60%	97.82%
Specificity	92.41%	96.37%
Recall	94.60%	97.82%
F1-Score	94.84%	97.63%
G-Mean	93.49%	97.36%
AUC	98.23%	99.56%

In fig.3 and table 2 results obtained are the comparison among different existing techniques like PCC with SVM vs. proposed techniques for the performance. The performance parameters are AUC, accuracy, recall, G-mean, F1-score, precision, sensitivity and specificity. On the basis of these results showed that for different parameters the performance values of proposed technique are higher as compared to others.

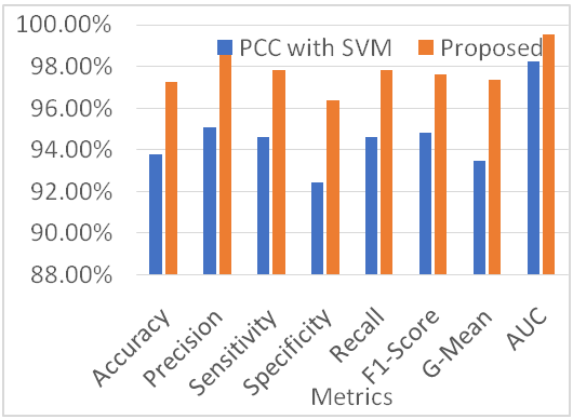

Fig. 3: Comparison results among existing method and proposed method analyzed on the basis of different parameters

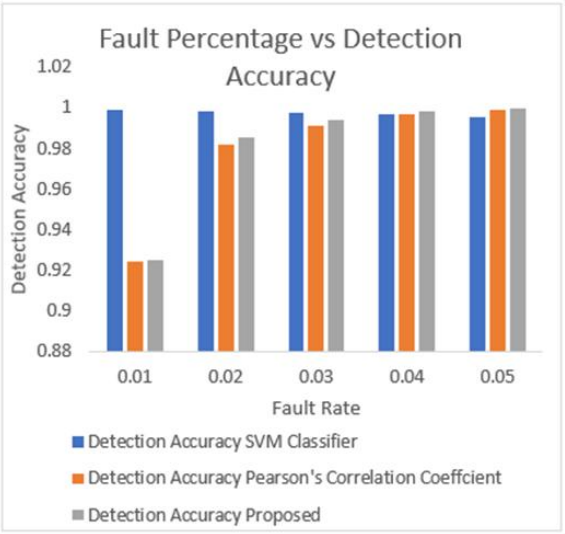

Fig. 4: Comparison results among existing method and proposed method analyzed on the basis of fault percentage and detection accuracy

Table 3: Comparison among existing methods and proposed method based on fault percentage vs. detection accuracy

Fault percentage	Detection Accuracy		
	SVM Classifier	Pearson's Correlation Coefficient	Proposed
0.01	0.999	0.924	0.925
0.02	0.998	0.982	0.985
0.03	0.9975	0.991	0.994
0.04	0.9965	0.997	0.998

0.05	0.995	0.999	0.9995

In fig. 4 and table 3 results obtained are the comparison among different existing techniques vs. proposed techniques for the performance. The performance parameters are Accuracy and fault rate measured in form of percentage. On the basis of these results showed that for different parameters the performance values of proposed technique are higher as compared to others. Below table 4 shows the comparison of existing and proposed method based on FPR.

Table 4: Comparison results among existing methods and proposed method based on fault positive rate

Methods	Average FPR (%)
SVM Classifier	1.93%
Pearson's Correlation Coefficient	1.65%
Proposed	0.96%

Table 5: Comparison results among existing methods and proposed method based on ROC curve

FPR	TPR		
	Proposed	Pearson's Correlation Coefficient	SVM Classifier
0	0	0	0
0.05	0.43	0.23	0.1
0.1	0.71	0.46	0.3
0.15	0.84	0.71	0.5
0.2	0.94	0.86	0.65
0.25	0.98	0.91	0.84
0.3	1	0.94	0.92
0.35	1	0.98	0.94
0.4	1	1	0.97
0.45	1	1	0.99
0.5	1	1	1
0.55	1	1	1
0.6	1	1	1
0.65	1	1	1
0.7	1	1	1
0.75	1	1	1
0.8	1	1	1
0.85	1	1	1
0.9	1	1	1
0.95	1	1	1
1	1	1	1

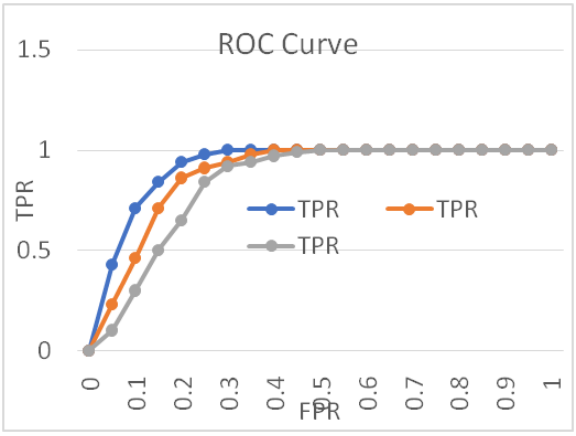

Fig. 5: Comparison results among existing methods and proposed method based on ROC curve

In fig.5 and table 5 results obtained are the comparison of existing vs. proposed techniques for the performance value. The performance parameters are TPR vs FPR to formed the ROC curve. On the basis of this graph results showed that for different parameters the performance of proposed technique is higher as compared to others.

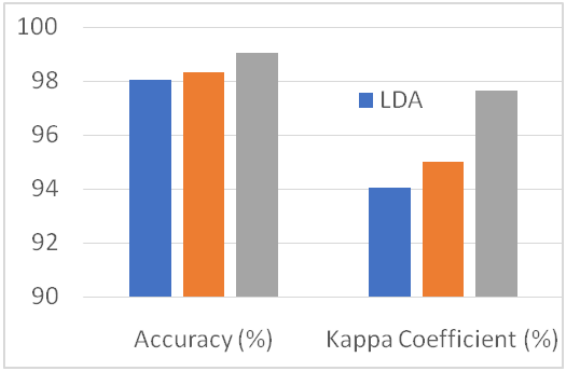

Fig. 6: Comparison results among existing methods and proposed method based on Kappa coefficient and accuracy

Table 6: Comparison results among existing methods and proposed method based on Kappa coefficient and accuracy

Classification models	Accuracy (%)	Kappa Coefficient (%)
LDA	98.03	94.06
CART	98.31	94.99
Proposed	99.02	97.64

In fig.6 and table 6 results obtained are among the existing methods and proposed method based on Kappa coefficient and accuracy. On the basis of these results showed that for different parameters the performance of proposed technique is better as compared to other.

CONCLUSION

The proposed optimization method is implemented to select best attribute among the overall feature database. This is done by using the Recurrent Distribution Correlated Optimization (RDCO) algorithm to find the relevancy between the feature of query data and from the entire dataset and selects the best optimal features. In this, the data is preprocessed and clustered by using the Maximum Possibility Combination (MPC) based clustering algorithm. To find the matching feature, the Multi-Block Convoluted Learning (MBCL). Finally, the proposed classification process of dataset can be compared with 70%, 80% and 90% training data. The proposed method can be implemented in the Python scripting and validate the performance by using the parameters like, detection rate, accuracy, fault percentage, false positive rate and true positive rate with the reference of concerned database.

Several issues with time and cost comparisons and calculations will need to be addressed in future work.

ACKNOWLEDGMENT

A heartfelt thanks is conveyed to Professors, North East Frontier Technical University, Medog, Aalo, Arunachal Pradesh, India., by authors for providing them the mandatory facilities to complete the project effectively.

REFERENCES

[1] Alsuwaiket, M., Blasi, A.H. and Al-Msie'deen, R.F. (2020). Formulating module assessment for improved academic performance predictability in higher education. Engineering, Technology & Applied Science Research.,9(3):4287- 4291.

[2] Rao, K.S., Swapna, N. and Kumar, P.P. (2018). Educational data mining for student placement prediction using machine learning algorithms. International Journal Engineering Technological Sciences., 7(12):43-46.

[3] Bharara, S., Sabitha, S. and Bansal, A. (2018). Application of learning analytics using clustering data Mining for Students' disposition analysis. Education and Information Technologies., 23(2): 957-984.

[4] X. Zhou, X. Liang, X. Du, and J. Zhao, "Structure based user identification across social networks," IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 6, pp. 1178–1191, 2018.

[5] P. H. Son, "A novel automatic picture fuzzy clustering method based on particle swarm optimization and picture composite cardinality," Knowledge-Based Systems, vol. 109, pp. 48–60, 2016.

[6] Zaffar, M., Savita, K.S., Hashmani, M.A. and Rizvi, S.S.H. (2018). A study of feature selection algorithms for predicting students' academic performance. International Journal of Advanced Computer Science and Applications., 9(5):541-549.

[7] J. Albericio, P. Judd, T. Hetherington, T. Aamodt, N. E. Jerger, and A. Moshovos, "Cnvlutin," ACM SIGARCH - Computer Architecture News, vol. 44, no. 3, pp. 1–13, 2016.

[8] S. Han, X. Liu, H. Mao et al., "Eie," ACM SIGARCH - Computer Architecture News, vol. 44, no. 3, pp. 243–254, 2016

[9] F. Wang, V. K. Devabhaktuni, and C. Xi, "Neural network structures and training algorithms for RF and microwave applications," International Journal of RF and Microwave Computer-Aided Engineering, vol. 9, no. 3, pp. 216–240, 2015

[10] Sokkhey, P. and Okazaki, T. (2020). Developing Web-based Support Systems for Predicting Poor-performing Students using Educational Data Mining Techniques. International Journal of Advanced Computer Science and Applications., 11(7):23-32.

[11] Utari, M., Warsito, B. and Kusumaningrum, R. Implementation of Data Mining for Drop-Out Prediction using Random Forest Method. International Conference on Information and Communication Technology., pp. 1-5, 2020.

[12] RİMİ, A.A., IBRAHİM, A.A. and BAYAT, O. (2020). Developing Classifier for the Prediction of Students' Performance Using Data Mining Classification Techniques. AURUM Mühendislik Sistemleri ve Mimarlık Dergisi., 4(1):73-91.

[13] Ashraf, M., Zaman, M. and Ahmed, M. An Intelligent Prediction System for Educational Data Mining Based on Ensemble and Filtering approaches. International Conference on Computational Intelligence and Data Science., 167, pp.1471-1483,2020.

[14] Adekitan, A.I. and Salau, O. (2019). The impact of engineering students' performance in the first three years on their graduation result using educational data mining. Heliyon., 5(2): e01250.

[15] X. Zhang, Y. Tian, and R. Cheng, "A decision variable clustering-based evolutionary algorithm for large-scale many objective optimizations," IEEE Transactions on Evolutionary Computation, vol. 99, 2016.

[16] Z. Lv, Y. Han, A. K. Singh et al., "Trustworthiness in industrial IoT systems based on artificial intelligence," IEEE Transactions on Industrial Informatics, vol. 99, p. 1, 2020.

[17] A. Zl, A. Dc, A. Rl, and B. Aa, "Artificial intelligence for securing industrial-based cyber–physical systems," Future Generation Computer Systems, vol. 117, pp. 291–298, 2021.

[18] T. ˙Inkaya, S. Kayalıgil, and N. E. Ozdemirel, "Ant Colony ¨ Optimization based clustering methodology," Applied Soft Computing, vol. 28, pp. 301–311, 2015.

[19] X. Xu, D. Cao, Y. Zhou, and J. Gao, "Application of neural network algorithm in fault diagnosis of mechanical intelligence," Mechanical Systems and Signal Processing, vol. 141, Article ID 106625, 2020.

[20] P. H. Son, "A novel automatic picture fuzzy clustering method based on particle swarm optimization and picture composite cardinality," Knowledge-Based Systems, vol. 109, pp. 48–60, 201

[21] S. A. Rahman, H. Tout, C. Talhi, and A. Mourad, "Internet of things intrusion detection: Centralized, on-device, or federated learning?" IEEE Network, vol. 34, no. 6, pp. 310–317, 2020.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] T. Ryffel, A. Trask, M. Dahl, B. Wagner, J. Mancuso, D. Rueckert, and J. Passerat - Palmbach, "A generic framework for privacy preserving deep learning," arXiv preprint arXiv:1811.04017, 2018.

[24] Y. Liu, S. Garg, J. Nie, Y. Zhang, Z. Xiong, J. Kang, and M. S. Hossain, "Deep anomaly detection for time-series data in industrial iot: A communication-efficient on-device federated learning approach," IEEE Internet of Things Journal, pp. 1–1, 2020.