Medical Data Classification Based on SMOTE and Recurrent Neural Network

P. Penchala Prasad, F. Sagayaraj Francis, S. Zahoor-Ul-Huq

Abstract*: Medical data classification analysis the medical data of the patients to predict the diseases risk. Data mining techniques were highly used in the medical data classification and predicted the diseases. Many existing methods were use the various classifier and feature selection to improve the performance of the classification. Although data imbalance problem is need to be solved for increases the performance. In this research, Synthetic Minority Over-sampling TEchnique (SMOTE) techniques is used for solving the data imbalance problem and Recurrent Neural Network (RNN) was used for the classification. The SMOTE method based on the k Nearest Neighbor (kNN) for the over-sample and under-sample the attributes. The RNN process the instance independent of the previous instance for the classification. Four medical datasets of University of California, Irvine (UCI) were used to evaluate the effectiveness of the proposed SMOTE-RNN method. The proposed SMOTE-RNN method has the accuracy of 85 % while existing method has 82 % accuracy.

Keywords: Data Imbalance, k Nearest Neighbor, Medical data classification, Recurrent Neural Network and Synthetic Minority Over-sampling Technique.

I. INTRODUCTION

Healthcare data classification is much useful in disease diagnosis and treating the patient in the early stage. The Healthcare classification method acts as a Decision Support System to improve the healthcare quality [1]. Medical datasets consist of more missing data and this affects the prediction performance. Hence, the data imputation method often used in the medical dataset and statistical methods were used to solve this problem. Many data imputation methods were applied in the existing methods and process the prediction [2]. Computational Intelligences such as Artificial Neural Network (ANN), fuzzy logic and optimization methods like Genetic Algorithm (GA), Particle Swarm Optimization (PSO) to increase the performance of the prediction [3]. In medical data, several methods were applied to increases the classification accuracy of the diseases. Network has been trained with the sufficient information to increase the prediction accuracy [4]. Many attributes present in the data affects the performance of the prediction and also same with the low number of attributes. Attribute selection plays major role in the prediction of diseases and hence, it is important to use the efficient classifier [5].

For instance, well-known classifier such as Naïve Bayes classify a new data point based on the probabilities processed in the training phase [6]. The Ada Boost and the Support Vector Machine (SVM) are the intelligent classifier that effectively handle the two-class task. These methods convert the multi-class problems into one or two class problems [7]. Healthcare data classification received the great attentions among the researchers in data mining area. Recently, a Single Layer Perceptron classifier is one of most popular data mining technique, which has been applied in healthcare data classification [8]. Many existing methods involves in applying the various classification technique for effective prediction of diseases. These methods still didn't solve the problem of the data imputation method in medical data classification [9, 10]. In order to solve the problem of data imputation, the SMOTE method is used with RNN to increase the performance of medical data classification. The SMOTE method uses the kNN to sample the medical data. The RNN techniques uses the attribute to predict the disease.

The paper is formulated as Literature review of recent medical data classification research were given in the section 2, Proposed SMOTE-RNN method is described in the section 3, experimental setup is provided in the section 4 and experimental result analysis is provided in the section 5 and conclusion of the research is given in section 6.

II. LITERATURE SURVEY

Data mining techniques were used in the medical data to increase the medical diagnostic for prediction of disease. Several data mining techniques were applied in the large amount of medical data for accurate prediction of diseases. Recent researches involves in the prediction of the diseases is surveyed in this section.

Yang, et al. [11] provided the technique for improve the classic method of Iterative Dichotomiser 3 (ID3) algorithm. The method involves in apply the balance function for the test attribute selection and heuristic method for the classification. The UCI medical data had been used to analyze the developed method performance. The developed method is compared with the traditional ID3 method, Decision tree and Random forest. The improved id3 method has proved to solve the problem of multi-value bias problem in attribute selection and numeric attribute discretization.

Revised Manuscript Received on February 05, 2020.
***** Correspondence Author
P. Penchala Prasad*, CSE department, G.PullaReddyEngineering College , Kurnool, India. Email: prasad.cse@gprec.ac.in
Dr. F. Sagaraj Francis, CSE department, Pondicherry Engineering College , Puducherry, India. Email: fsfrancis@pec.edu
Dr. S. Zahoor-Ul-Huq, CSEdepartment, G. Pulla Reddy Engineering College , Kurnool, India.. Email: szahoor@gmail.com

The experimental evaluation shows that the developed method has the higher performance in disease prediction than traditional id3 and other classifiers. This method need to solve the problem of imbalance dataset, which affects the performance of the developed method.

Bania and Halder [12], proposed an ensemble attribute selection based on the concepts of rough set theory, which is named as R-ensemble. The developed R-ensemble method analysis the attribute class, attribute relevance and its importance to select the subset of attributes. The developed R-ensemble classifier involves in analysis the attributes and combining multiple attributes subsets based on the different rough set filter to select the optimal attribute subsets in the classification. The input dataset is preprocessed based on the k-Nearest Neighbors (kNN) method to solve the problem of impute data in the dataset. The UCI medical datasets were used to evaluate the effectiveness of the proposed R-ensemble method. The proposed R-ensemble method is tested with three classifiers and this shows that the developed method has the higher performance in disease prediction by proper selection of attributes. The developed method efficiency is affected by the imbalance dataset in the function.

Chen, et al. [13] explore the performance of the Extreme Learning Machine (ELM) and Kernel ELM (KELM) method in the prediction of Parkinson's disease. The various parameters including kernel parameter, type of activation function in ELM and number of hidden neuron and constant parameter were analyzed in detailed. Four feature selection methods were applied to improve the effectiveness of the classifier. The experimental result demonstrate that the developed method has the higher performance than existing method. The feature selection techniques are not efficient and imbalance dataset is need to be solved.

Shen, et al. [14] developed fruit fly optimization algorithm for the parameter optimization in the Support Vector Machine (SVM) for disease prediction. The SVM parameters were effectively analyzed by the optimization algorithm and provide the optimal parameter set. The UCI medical datasets were used to evaluate the performance of the developed method. The various existing optimization methods were applied in SVM parameter optimization and compared with existing methods. The experimental analysis shows that the proposed fruit fly optimization technique has the higher efficiency than optimization technique. The imbalance in the datasets were need to be treated for effective prediction of diseases.

Alam, et al. [15] developed features ranking method for the Rand forest classifier for the prediction of the disease in the medical data. Several ranking methods were used to rank the features and features with the highest rank were used for the classification. The UCI medical datasets were used to evaluate the effectiveness of the developed method. The experimental result shows that the developed method was applied to provide various datasets with more performance. The data imbalance problems were not solved in this research, which affects the effectiveness of the developed model.

From the analysis of the various recent researches in the medical data classification/disease prediction, the researchers were focused on the attribute selection or ensemble method that provides the higher performance. Although data imbalance problem is neglected in these researches, which affects the performance of the developed method. In order to solve this problem, the SMOTE method is proposed with RNN.

III. PROPOSED METHOD

Medical data classification is the process of predicting the diseases using the attributes of the medical data. Many existing methods applies the different classification and feature selection techniques to predict the diseases. In this research, SMOTE techniques were applied for the solving the data imbalance problem by over-sampling the input data and RNN technique is used for the feature selection and predictions of diseases. The UCI medical data were used to evaluate the developed method effectiveness. The experimental results show that the developed method has the higher effectiveness in the medical data classification. The overview of the developed SMOTE-RNN method is shown in the block diagram of Fig. (1). The detailed description of the SMOTE and RNN were given in this section. In the SMOTE method, the classes with the minimum attributes were applied with over-sample and the classes with higher attributes are treated with under-sampling. The samples that were applied to the RNN are consider as independent to each other and classify the data.

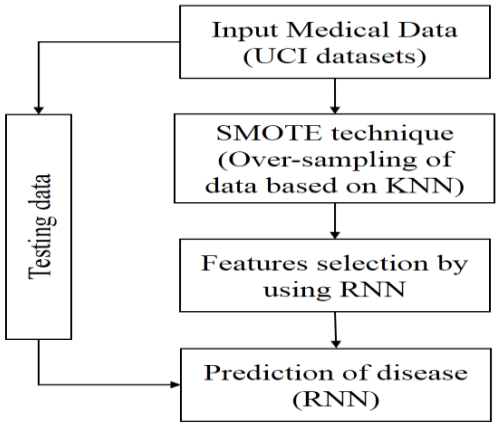

Fig. 1.The overview of the proposed SMOTE-RNN method for medical data classification

A. Synthetic Minority Over-sampling TEchnique

The research develops an over-sampling method in which the minority class is over-sampled by creating "synthetic" example instead of replacement. This method is inspired from the research that has achieves the higher efficiency in the handwritten character recognition. Extra training data has been created in that method by performing some operation on real data. The research considers the rotation and skew were the natural way to synthetic the training data. This method generates the synthetic data in a less application-specific manner and this method process in the "feature space" rather than "data space". Each minority sample is over-sampled and synthetic example is provided along the line segments based on the k minority class nearest neighbors.

The neighbors are selected from the k Nearest Neighbor (KNN) depend on the number of over-sampled required. The KNN method selects the neighbor randomly [16]. In this research, there are 5 classes is set for KNN to select the neighbors. For example, if the amount of over-samples required is 200%, then two neighbors are selected from the 5 classes. One sample is generated at the direction of each neighbor. Synthetic samples are generated as follows: Calculate the difference between feature vector under consideration and its nearest neighbor, multiply this difference with random number between 0 and 1, and add the feature vector. The SMOTE algorithm is briefly discussed in the research [16]. The sample input features were applied to the RNN, where attributes were selected and disease are predicted. The detailed descriptions of the RNN are explained in the next sub-section.

B. Recurrent Neural Network

The RNN provides the effective solution in the problem of the dynamic time series analysis. These method is high used in the Natural Language Processing, Handwritten recognition and speech recognition task [17]. The RNN is trained with time series vector sequence $X_{t-1}, X_t, X_{t+1}\ldots$. The hidden layer S_t is develops based on input X_t and the previous hidden layer S_{t-1}. The RNN process is described in the following Equations.

$$S_t = f\left(U.X_t + W.S_{t-1}\right) \quad (1)$$

$$O_t = g\left(V.S_t\right) \quad (2)$$

Where, S_t denotes the memory sample at time t, the value of the hidden layer is calculated based on Eq. (1). The previous moment output is denoted as W and this is used as the weighted input at this moment, and input sample weight is denoted as U. The output value O_t is measured based on the Eq. (2), with V is the sample output weight. The activation functions are f and g, where f is the Sigmoid, ReLU or tanh activation function and g is generally softmax activation function.

As the structure of RNN grows, the gradient measured using the hidden layer back propagation may vanish or explode. The gradient cropping can be used to solve the explosion problem, but this doesn't solve the vanishing problem. The RNN cannot easily capture the dependence in the large distance of sequence model. The Long Short Term Memory LSTM can be used to solve this vanishing problem. The core of LSTM is the state cell or cell state. This consists of three types of gates: the input, output and forget gate. The related formulas are show in Eq. (3-7).

$$f_t = \sigma\left(W_f.\left[h_{t-1}, x_t\right] + b_f\right) \quad (3)$$

$$i_t = \sigma\left(W_i.\left[h_{t-1}, x_t\right] + b_o\right) \quad (4)$$

$$o_t = \sigma\left(W_o.\left[h_{t-1}, x_t\right] + b_o\right) \quad (5)$$

$$C_t = f_t * C_{t-1} + i_t * \tanh\left(W_C.\left[h_{t-1}, x_t\right] + b_c\right) \quad (6)$$

$$h_t = o_t * \tanh\left(C_t\right) \quad (7)$$

The three multiplicative gates are given in the Eq. (3-5) namely: forget gate, input and output gate. The $\left[x_t, h_{t-1}\right]$ in Eq. (3) is input and the activation function is denoted by σ, the parameters are different. The C_t denotes the cell state in Eq. (6-7), which is obtained from previous input time step C_{t-1}. If the forget gate is obtained as 0, then the previous state moment is completely cleared to consider the input at time step. The input is received or not is determined by the input gate i_t and likewise, whether to provide output or not is decided by final output gate o_t. The performance analysis of the SMOTE-RNN in medical data classification were evaluated in the next section.

IV. EXPERIMENTAL SETUP

The experimental setup of the developed SMOTE-RNN and parameter settings were descripted in this section. The UCI medical datasets were used in this method to evaluate the developed method effectiveness. The experiment was carried out in the system consists of Intel i7 processor with 8 GB of RAM and 500 GB hard disk.

A. Datasets

Four UCI medical datasets were applied to evaluate the effectiveness of the developed method and the existing methods. The details of the datasets were given in the table 1 with number of attributes and classes.

Table- I: The UCI datasets details

Name of dataset	Samples	Attributes	Classes
Wisconsin breast cancer	699	10	1
Lung cancer	32	56	2
Mammographic mass	961	6	1
Statlog (Heart)	270	13	1

B. Evaluation Metrics

The three important metrics such as accuracy, Root Mean Square Error (RMSE), and F-measure were measured from the performance of the proposed method in the medical data classification. The accuracy metric provides the correctness of the classification, RMSE provides the predictive error of the model and F-measure provides the mean value of the precision and recall. The formula for the accuracy, RMSE and F-measure were provided in the Eq. (8-10).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}\left(\widehat{y_t} - y_t\right)^2}{T}} \quad (9)$$

$$F - measure = \frac{2TP}{2TP + FP + FN} \quad (10)$$

Where TP denotes the True positive, FN denotes the False Negative, FP denotes the False Positive and TN denotes the True Negative.

C. Parameter settings

In SMOTE, the kNN is used for the sampling and five nearest neighbors used in the method. In the RNN parameter settings, Adam is used to train the network and parameter settings at

$\beta_1 = 0.9, \beta_2 = 0.999 and \in = 10e^{-8}$. The 10-fold cross-validation is applied to investigate the effectiveness of the developed method. The experimental results of the proposed SMOTE-RNN were explained in the next section.

V. EXPERIMENTAL RESULTS

Medical data classification helps to predict the disease early and various existing methods were proposed in the medical data classification. Most of the existing methods were suffer from the data imbalance problem. This research involves in applying the SMOTE-RNN method for efficient prediction of diseases. There are four UCI datasets were used to evaluate the proposed SMOTE-RNN method effectiveness in medical data classification, which were described in this section.

A. Wisconsin dataset

Wisconsin dataset was used to analyze the developed method performance and compared with existing methods. Accuracy and RMSE metrics were calculated to analyze the effectiveness and error value.

Table- II: The various classification method in Wisconsin dataset

Methods	Accuracy	RMSE
SMOTE-RNN	0.9617	0.2032
Id3' [11]	0.9528	0.2197
Id3	0.9056	0.2058
J48	0.9442	0.218
Decision tree	0.9242	0.2587
Random Forest	0.9285	0.2675

The Wisconsin dataset were used to investigate the effectiveness of the proposed SMOTE-RNN method, as shown in the Table 2. The table shows that the SMOTE-RNN method has the higher efficiency, due to the data imbalance is solved in the method. Along with the RNN has the advantage of process the instance with the independence of previous instance. This helps to increase the performance of the developed SMOTE-RNN method.

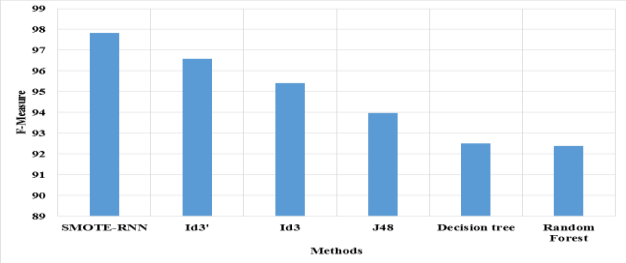

Fig. 2.Wisconsin dataset F-measure

The various methods were compared with proposed SMOTE-RNN method in the Figure 2. The F-measure was evaluated from various methods and compared with each other in the Figure 2. The figure shows that the developed SMOTE-RNN method has the higher efficiency than other existing methods. The SMOTE-RNN method has the higher F-measure due to data imbalance is reduced in the method.

B. Lung cancer dataset

Lung cancer dataset was used in this to evaluate the performance of developed and existing methods. Two metrics such as accuracy and RMSE were calculated and analyzed with various existing methods.

Table- III: Lung cancer dataset analysis

	Accuracy	RMSE
SMOTE-RNN	0.5621	0.5348
Id3' [11]	0.5313	0.559
Id3	0.5	0.5578
J48	0.5	0.5005
Decision tree	0.4375	0.4797
Random Forest	0.4375	0.5848

The accuracy and RMSE is measured for the various methods in lung cancer dataset, as shown in Table 3. The table shows that the developed SMOTE-RNN method has the higher effectiveness than other existing method in the medical data classification. The data imbalance is reduced in the proposed method that helps to increase the efficiency of the method. Moreover, the RNN process the instance with independent to the previous instance. These two advantage helps to increase the performance of the proposed SMOTE-RNN method.

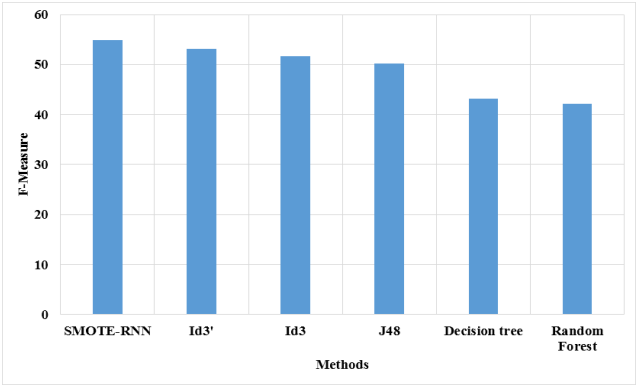

Fig. 3.F-measure on Lung dataset

The F-measure for the proposed SMOTE-RNN and other existing methods in the medical data classification is compared in the Figure 3. The proposed SMOTE-RNN method has the higher efficiency than other existing methods due to the reduction of data imbalance in the method. Existing method has the lower performance due to the various number of attributes present for the classes.

C. Mammographic mass dataset

Mammographic mass dataset is UCI medical dataset used to evaluate the effectiveness of the developed method. The error value is analyzed in the classification and analyzed with the existing classification method.

Table- IV: Performance measure in Mammographic mass dataset

	Accuracy	RMSE
SMOTE-RNN	0.854	0.1673
Id3' [11]	0.8262	0.1738
Id3	0.6545	0.1977
J48	0.821	0.179
Decision tree	0.8189	0.1811
Random Forest	0.7659	0.2341

The accuracy and RMSE of the proposed and existing method is measured in the Mammographic mass dataset, as provided in the Table 4. The table shows that the SMOTE-RNN method has the higher accuracy and lower error in the performance due to the advantages of imbalance reduction and instance is processed independently. The proposed SMOTE-RNN method

has the accuracy of 85 % in the mammographic mass dataset.

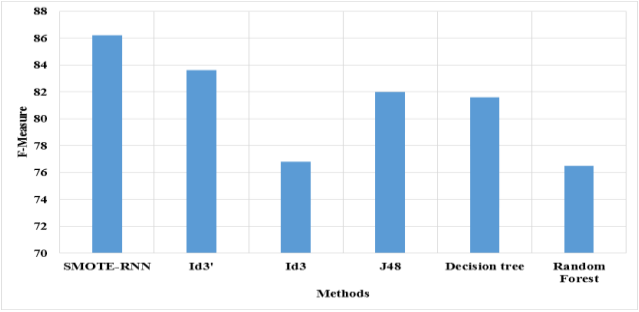

Fig. 4.F-measure on mammographic dataset

The proposed SMOTE-RNN and existing methods is measured with F-measure and compared in the Figure 4. The proposed SMOTE-RNN method has the higher performance due to the advantages of reduction of imbalance data and process the instance independently. The proposed SMOTE-RNN method has the higher F-measure of 86% and existing method has 83 % of accuracy.

D. Statlog dataset

The statlog dataset were used in this research for the investigate the efficiency of the developed method.

Table- V: Performance analysis on Statlog dataset

	Accuracy	RMSE
SMOTE-RNN	0.7892	0.4342
Id3' [11]	0.7778	0.4699
Id3	0.3519	0.602
J48	0.7667	0.4601
Decision tree	0.7259	0.4477
Random Forest	0.7407	0.5092

The proposed SMOTE-RNN and existing methods were investigated in the statlog dataset, as shown in the Table 5. The proposed SMOTE-RNN has the higher performance due to the data imbalance reduction and the instance is processed independently. The accuracy of the proposed SMOTE-RNN is 78%, while existing method has 77%.

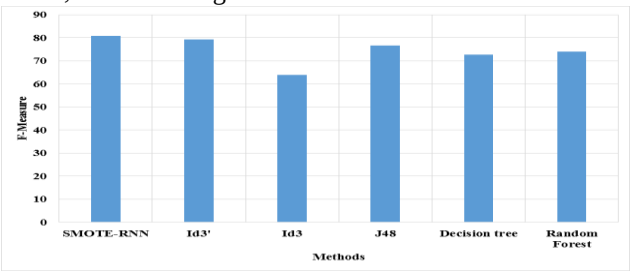

Fig. 5.F-measure on Statlog dataset

The proposed and existing method is calculated with F-measure and compared in the Figure 5. The figure shows that the developed SMOTE-RNN method has the higher F-measure due to the reduction of data imbalance in the method. The SMOTE-RNN has the F-measure of 80 %, while Id3' method has 79%.

Therefore, the proposed SMOTE-RNN method has the higher efficiency than the existing method due to the two advantages such as reduction of data imbalance and instances is process independently.

VI. CONCLUSION

Various data mining techniques were applied in the medical data classification to increase efficiency in the prediction. Many existing methods were suffering from the data imbalance problem, which is nature of the medical dataset. This research aims to apply the SMOTE-RNN method to solve the data imbalance problem and increase the predication performance. The SMOTE method sample the medical data based on the kNN method. The RNN selects the attributes for the classification and predict the diseases based on the training data. The RNN has the advantage of process the instances of data with independent to previous instances. Four UCI medical datasets were used to analyze the performance of the SMOTE-RNN method and compare with the existing methods. The SMOTE-RNN method has achieved the accuracy of 85% in mammography mass dataset, while existing methods has 82 % accuracy. The proposed SMOTE-RNN method has the higher performance than the Id3 and decision tree method. The future work of the developed method involves in hybrid attribute selection to increase the efficiency of the medical data classification.

REFERENCES

1. T. Nguyen, A. Khosravi, D. Creighton, and S. Nahavandi. (2015). Medical data classification using interval type-2 fuzzy logic system and wavelets. Applied Soft Computing, 30, pp. 812-822.
2. U. Yelipe, S. Porika, and M. Golla. (2018). An efficient approach for imputation and classification of medical data values using class-based clustering of medical records. Computers & Electrical Engineering, 66. pp. 487-504.
3. B. Dennis, and S. Muthukrishnan. (2014). AGFS: Adaptive Genetic Fuzzy System for medical data classification. Applied Soft Computing, 25, pp. 242-252.
4. C. Y. Fan, P. C. Chang, J. J. Lin, and J. C. Hsieh. (2011). A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. *Applied Soft Computing, 11(1)*, pp. 632-644.
5. D. C. Li, C. W. Liu, and S. C. Hu, (2011). A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets. *Artificial Intelligence in Medicine, 52(1)*, pp. 45-52.
6. A. Wood, V. Shpilrain, K. Najarian, and D. Kahrobaei. (2019). Private naive bayes classification of personal biomedical data: Application in cancer data analysis. *Computers in biology and medicine*, *105*, pp.144-150.
7. C. Y. Fan, P. C. Chang, J. J. Lin, and J. C. Hsieh. (2011). A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. *Applied Soft Computing, 11(1)*, pp. 632-644.
8. C. Zhang, L. Zhu, C. Xu, and R. Lu. (2018). PPDP: An efficient and privacy-preserving disease prediction scheme in cloud-based e-Healthcare system. *Future Generation Computer Systems*, *79*, pp. 16-25.
9. O. Graa, and I. Rekik. (2019). Multi-view learning-based data proliferator for boosting classification using highly imbalanced classes. *Journal of neuroscience methods*, *327*, pp. 108344.
10. R. J. Kuo, P. Y. Su, F. E. Zulvia, and C. C. Lin. (2018). Integrating cluster analysis with granular computing for imbalanced data classification problem–A case study on prostate cancer prognosis. *Computers & Industrial Engineering, 125*, pp. 319-332.
11. S. Yang, J. Z. Guo, and J. W. Jin. (2018). An improved Id3 algorithm for medical data classification. *Computers & Electrical Engineering, 65*, pp. 474-487.
12. R. K. Bania, and A. Halder. (2020). R-Ensembler: A greedy rough set based ensemble attribute selection algorithm with kNN imputation for classification of medical data. *Computer Methods and Programs in Biomedicine*, 184, pp. 105122.
13. H. L. Chen, G. Wang, C. Ma, Z. N. Cai, W. B. Liu, and S. J. Wang. (2016). An efficient hybrid kernel extreme learning machine approach for early diagnosis of Parkinson' s disease. *Neurocomputing, 184*, pp.131-144.
14. L. Shen, H. Chen, Z. Yu, W. Kang, B. Zhang, H. Li, B. Yang, and D. Liu. (2016). Evolving support vector machines using fruit fly optimization for medical data classification. *Knowledge-Based Systems*, *96*, pp. 61-75.

15. M. Z. Alam, M. S. Rahman, and M. S. Rahman. (2019). A Random Forest based predictor for medical data classification using feature ranking. *Informatics in Medicine Unlocked*, *15*, pp. 100180.
16. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, pp. 321-357.
17. T. Li, Z. Zhang, and H. Chen. (2019). Predicting the combustion state of rotary kilns using a Convolutional Recurrent Neural Network. *Journal of Process Control*, *84*, pp. 207-214.

AUTHORS PROFILE

Mr.P. Penchala Prasad, received his M.Tech degree from JNTUA University, Anantapur, Andhra Pradesh, India and pursuing Ph.D. in Department of Computer Science and Engineering at Pondicherry Engineering College, Pondicherry, India. He is presently working as Assistant Professor in the Department of Computer Science and Engineering at G. Pulla Reddy Engineering College (Autonomous): Kurnool, Andhra Pradesh, India. He has published in 4 journals.

Dr.F.Sagayaraj Francis, received his Ph.D., in computer Science and Engineering from Pondicherry University, Pondicherry India. He is presently working as professor in Department of Computer Science and Engineering at Pondicherry Engineering College, Pondicherry, India. His research interest includes Database Management Systems, Data Analytics, Geographical Information Systems, Big Data, and Information Systems. He has published in 30 journals and more than 20 International Conferences.

Dr. S. Zahoor Ul Huq obtained his M.E. degree from Anna University, Chennai, Tamilnadu, India and his Ph.D. in the field of Networks from Sri Krishna Devaraya University, Anantapur, Andhra Pradesh, India. He is presently working as Professor in the Department of Computer Science and Engineering at G. Pulla Reddy Engineering College (Autonomous):Kurnool, Andhra Pradesh, India and is also the Additional Controller of Examinations. He has presented Nineteen research papers at various International/ National Journals/Conferences. His major interest is in Networks, Databases and Object Oriented Programming.