Received: January 17, 2020. Revised: March 13, 2020. 10

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

The Prediction of Diseases using Rough Set Theory with Recurrent Neural Network in Big Data Analytics

Vamsidhar Talasila[1]* Kotakonda Madhubabu[1] Meghana Chakravarthy Mahadasyam[1] Naga Jyothi Atchala[1] Lakshmi Sowjanya Kande[1]

[1]*Koneru Lakshmaiah Education Foundation, Vaddeswaram, India*
* Corresponding author's Email: talasila.vamsi@kluniversity.in

Abstract: In a modern life, early healthcare prediction plays an important role to prevent the loss of life caused by prediction delays in treatment. Nowadays, the researchers focused on the Big data analysis, which is used to identify the future health status and provides an efficient way to overcome the issues in early prediction. Many researches are going on predictive analytics using machine learning techniques to provide a better decision making. Big data analysis provides great opportunities to predict future health status from health parameters and provide best outcomes. However, the data classification is one of the major challenging tasks due to noisy data or missing data in the dataset. Feature selection techniques play an important role in the classification process by removing irrelevant features from the extracted data. In this research work, the Rough Set Theory (RST) technique is used to select the most relevant features, which helps to provide the efficient classification of medical data and disease detection. The selected features are given as input to the Recurrent Neural Network (RNN) technique for disease prediction. The proposed method is also called as RST-RNN, where the experiments are carried out on the UCI machine learning repository dataset in terms of accuracy, f-measure, sensitivity and specificity. The results showed that the RST-RNN method achieved accuracy of 98.57%, where the existing Support Vector Machine (SVM) achieved 90.57% accuracy and Naive Bayes (NB) achieved 97.36% accuracy for heart disease dataset.

Keywords: Big data analysis, Decision making, Feature selection technique, Naive bayes, Rough set theory, Recurrent neural network.

1. Introduction

Over the past two decades, digital data is becoming increasingly important in many domains like healthcare, science, technology and society. A large amount of data has been captured and generated from multiple areas, multiple sources such as streaming machines, high throughput instruments, sensor networks, mobile application and especially in healthcare, this huge amount of collected data is represented as big data [1-2]. The process of storing, visualizing and extraction of knowledge through various huge data types has become a challenge, because of using inadequate existing technologies tools. One of the most important technological challenges of big data analytics is to identify efficient ways to obtain the valuable information for different types of users [3]. Currently, the various forms of healthcare data sources are being collected in both clinical and non-clinical environments, where the digital copy of a patient's medical history are the most important data in healthcare analytics. Therefore, designing a distributed data system to deal with big data faces three main challenges: First, it is difficult to collect data from distributed locations due to the heterogeneous and huge volume of data [4]. Second, storage is the main problem for heterogeneous and massive datasets. The last challenge is related to big data analytics, more precisely to mining massive datasets in real-time or near real-time that include modeling, visualization, prediction, and optimization [5]. These challenges require a new processing paradigm, as the current data management systems are not efficient in dealing

with the heterogeneous nature of data or the real-time [6].

Improper diagnosis may cause death or disability to the patient. Disease Prediction Model can support medical professionals and practitioners in predicting the particular disease. The huge amount of data that can be collected using digital devices (by the patient itself in hospital) can use with big data to diagnose patients and predict diseases [7-8]. But, applying machine learning on this big data stream is challenging as the traditional machine learning systems are not suitable to handle such a massive volume or varied velocity. The analytical data processing is considered as another major problem, because efficient data integration between systems are involved by performing richer analytical data processing. Most of the existing works involve machine learning, but in case of real-time applications, the machine learning techniques are insufficient to handle the big data [9-10] Classification techniques are widely used in healthcare, since they are capable of processing large set of data. The common used techniques in healthcare are NB, SVM, Nearest Neighbor (NN), decision tree (DT), Fuzzy logic, Fuzzy based neural network (FNN), Artificial neural network (ANN), and genetic algorithms (GA) [11]. Machine learning with classification can be efficiently applied in medical applications for complex measurements. Modern classification techniques provide more intelligent and effective prediction techniques for disease prediction [12]. In this research, the important features are selected by using RST method which is used to increase the performance of the classification technique. The important features from the medical dataset are given to RNN method for classifying the data. The overfitting is also reduced in the training data while using the RST method and the validation of the proposed RST-RNN method is conducted on various UCI datasets for predicting the diseases.

This research paper is prepared as follows, section 2 describes the review of existing techniques with its advantages and limitations. Section 3 explains the importance of proposed method for disease prediction. The experiments are conducted to validate the effective of a proposed RST-RNN method against existing techniques are presented in Section 4. Finally, the conclusion of the research work with future development is illustrated in section 5.

2. Literature review

In this section, discussions of existing techniques are presented, which are used to predict the health status of patients using machine learning techniques. In addition, the advantages and limitations of the existing methods were also discussed in [13-18].

R. Venkatesh, C. Balasubramanian, and M. Kaliappan, [13] designed a Big Data Prediction Analytics Model for heart disease prediction using NB technique as BPA-NB. This system used a probabilistic classification based on Bayes' theorem to analyse the data. To filter the unnecessary data, BPA-NB used the clustering technique and made the prediction in an effective way. The computational complexity was reduced by using MapReduce algorithm with Apache Spark framework. The experiments were carried out on UCI dataset to validate the effectiveness of BPA-NB against existing techniques by means of processing time, CPU utilization and accuracy. The BPA-NB method used for predicting only the heart disease and it provided poor performance on other diseases prediction.

M. Nilashi, O. Ibrahim, H. Ahmadi, L. Shahmoradi, and M. Farahmand, [14] developed the Incremental Support Vector Regression (ISVR) for Unified Parkinson's Disease Rating Scale (UPDRS). The prediction of Motor-UPDRS and Total-UPDRS was done by ISVR. In this method, a self-organizing map (SOM) was used to cluster the data and non-linear iterative partial least squares (NIPALS) for dimensionality reduction. To evaluate the ISVR method, several experimental analyses were conducted on a real-world PD dataset taken from UCI. The results indicated that the method that combines SOM, NIPALS, and ISVR techniques was effective in predicting the Total-UPDRS and Motor-UPDRS. The ISVR method reduced the computation time for only small data and also used some important attributes for PD diagnosis, where other attributes were not considered.

A. Di Noia, A. Martino, P. Montanari, and A. Rizzi, [15] predicted the occupational disease risks by using pattern recognition techniques and computational intelligence techniques namely SVM and KNN. A set of meaningful labelled clusters was determined as the final model by using k-means algorithm. The optimal hyper parameters and optimal ad-hoc dissimilarity measure weights were found out using genetic algorithms for classification systems and improved the performance of those systems. The experiments were carried out to estimate the performance of three techniques against existing technique on standard collected datasets by means of fitness functions for different classes. This

algorithm performed well only on occupational disease forecasting of the collected dataset.

L. R. Nair, D. Sujala Shetty, and D. Siddhanth Shetty, [16] aimed to develop a real-time remote health status prediction system on big data processing engine, Apache Spark, testing and deployed on a cloud, where DT was designed on streaming data. Through tweet streams, the relevant health data of user was received and then send a direct message to the user about their health status by using DT algorithm. A user received the information about his health status instantly and privately with a single tweet and which is used to decide whether he/she need expert health care or not. A variety of diseases were also used to predict by using slight modification of this DT algorithm. The recovery of data was not possible because tweets were deleted permanently after a certain time period.

T. Chen, J. Xu, H. Ying, X. Chen, R. Feng, X. Fang, and J. Wu, [17] predicted the Extubation Failure (EF) by analyzing 3636 adult patient records in MIMIC-III clinical database using Light Gradient Boosting Machine (LightGBM). According to the results of LightGBM, afeature importance analysis were carried out by interpreting these features using SHapley Additive exPlanations (SHAP). The experiments were carried out on the clinical database against existing techniques namely SVM, ANN and Logistic Regression (LR). The results stated that LightGBM achieved an accurate prediction than other existing techniques. However, the recognition for EF using LightGBM were still not very high.

F. Van Wyk, A. Khojandi, and R. Kamaleswaran, [18] presented the hierarchical analysis of machine learning algorithms for improving the predictions of at-risk patients. In addition, a multi-layer machine learning approach was developed to analyze the high-frequency and continuous data. The experimental results illustrated the capabilities of this approach for early identification of patients at risk of sepsis, where the physiological data collected from bedside monitors. By this analysis, the multi-layer machine learning approach potentially helped to reduce the mortality and morbidity in the ICU. Even though, the method having Sequential Organ Failure Assessment (SOFA) score, the onset of organ failure was not identified by using this algorithm.

N. Kausar, A. Abdullah, B. B. Samir, S. Palaniappan, B. S. AlGhamdi, and N. Dey, [19] implemented the hybrid approach namely SVM with K-means clustering technique for medical data classification. The attribute dimension was reduced by introducing the Principal Component Analysis (PCA) algorithm. Then, the related parameters and measures were adjusted effectively to differentiate the normal and abnormal patients. The experiments were carried out on the UCI datasets in terms of accuracy, precision, recall and f-measure. When the unseen patterns of similar behaviours were introduced within the selected clusters, the developed study reduced the detection rate with high classification time.

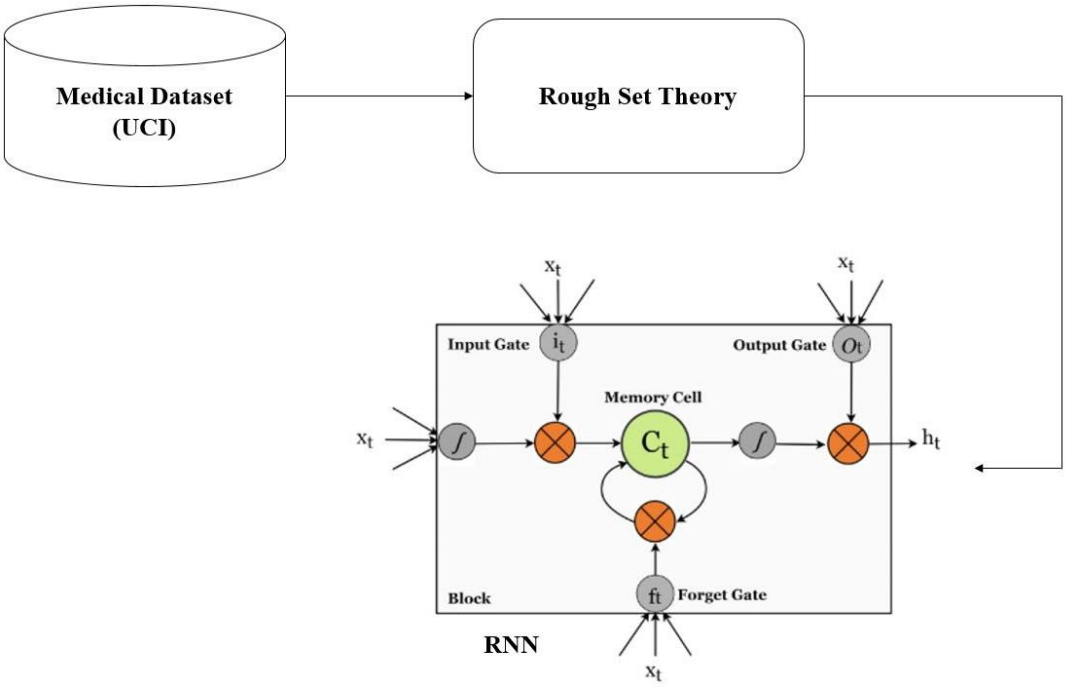

Figure. 1 The overview of the developed RST and RNN

The existing techniques are used to predict the disease either on only UCI dataset or collected dataset, where other diseases are not specified by these traditional techniques. In this research study, the RST-RNN method focused on five UCI datasets and selected the most important relevant features for better classification.

3. Proposed methodology

The disease prediction from healthcare data is a critical task due to the presence of a various relationship between the aspects of the patients and the disease. The disease prediction provides the many advantages like an early stage disease diagnosis and reduces the mortality rate. Healthcare data were present in the large amount and this need to be analyzed effectively. In this research, the RST and RNN is applied for disease prediction on medical data. The dependency between the attributes is found by analyzing the characteristics of attributes using RST and also used to remove the superfluous attributes. The RST generated decision rule was provided as input to the RNN. The RNN analyzes the attributes of the data with decision rule for disease prediction. This section will discuss the detailed information about the working of RST and RNN. The block diagram of RST and RNN in disease prediction is shown in Fig. 1.

3.1 Rough set theory

Let $I = (U, A)$ be an information system, where U is a nonempty set of finite objects called the universe of discourse; A is a non-empty set of attributes. With every attribute $a \in A$, a set of its values (V_a) is associated. For a subset of attributes $P \subseteq A$ there is an associated equivalence relation IND (P), which is called an indiscernibility relation. The relation IND (P) can be defined in following Eq. (1):

$$IND(P) = \{(x, y) \in U^2 | \forall a \in P, a(x) = a(y)\} \quad (1)$$

If $(x, y) \in IND(P)$, then x and y are indiscernible by attributes from P. The equivalence classes of the P-indiscernibility relation are denoted $[x]p$. The indiscernibility relation is the mathematical basis of the RST. The lower and upper approximations are two basic operations in RST. For a subset, $X \subseteq U$. X can be approximated using only information contained within P by constructing the P-lower approximation donated as $\underline{P}X$, is the set of all elements of U, which can be certainly classified as elements of X based on the attribute set P. The P-upper approximation of X, denoted as $\overline{P}X$, which can be possibly classified as elements of X based on the attribute set P. These two definitions are expressed in Eq. (2) & (3).

$$\underline{P}X = \{X | [X]_P \subseteq \underline{P}X\} \quad (2)$$

$$\overline{P}X = \{X | [X]_P \cap X \neq \varphi\} \quad (3)$$

Where, $\underline{P}X$ is illustrated as P-lower approximation and $\overline{P}X$ is depicted as P-upper approximation. The RST selects the features with dependency of attributes and reduces the superfluous features. The features selected by the RST are provided as input to the RNN for the classification.

3.2 Recurrent neural network

RNNs are a good solution to the problem of modeling dynamic changes in a time series. They are widely used in natural language processing, speech recognition, and handwriting recognition tasks. The RNN inputs the time change vector sequence X_{t-1}, X_t, X_{t+1}.. . As the sequence continues to advance, the hidden layer S_t is simultaneously affected by the input X_t, and the previous hidden layer S_{t-1}. The following Eq. (4), & (5) can be used to formally describe the RNN process:

$$S_t = f(U.X_t + W.S_{t-1}) \quad (4)$$

$$O_t = g(V.S_t) \quad (5)$$

Where, S_t represents the memory of the sample at time, t, i.e. the value of the hidden layer, as calculated by Eq. (4). W is the output of the previous moment, which is used as the weight input at this moment, and U is the sample weight of the input. The Eq. (5) is used to calculate the output value as O_t with V that describes the sample weight of the output. Both f and g are activation functions, where f can be an activation function such as tanh, ReLU, or the sigmoid. g is usually a softmax activation function.

As the RNN structure deepens, the gradient calculated by the hidden layer back propagation may vanish or explode. Although gradient cropping can cope with gradient explosions, but it failed solve gradient vanishing. So, in the text sequence of a language model, RNN cannot easily capture the dependence between the text elements across the

large distances in the sequence. The use of a long short-term memory (LSTM) can solve the aforementioned problems. The core of an LSTM is the state of the cell (i.e. cell state). It also includes three kinds of gate structure: the input, output and forget gate. Here, the relevant formulas Eq. (6-10) are as follows:

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \quad (6)$$

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \quad (7)$$

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \quad (8)$$

$$C_t = f_t \times C_{t-1} + i_t \times \tanh(W_c.[h_{t-1}, x_t] + b_C) \quad (9)$$

$$h_t = o_t \times \tanh(C_t) \quad (10)$$

Eq. (6), Eq. (7), Eq. (8) are three multiplicative gates: the forget gate, f_t; the input gate, i_t; and the output gate, o_t. The input in Eq. (6), Eq. (7), Eq. (8) is $[x_t, h_{t-1}]$, but the parameters are different. σ represents the sigmoid activation function. C_t in Eq. (9) is the cell state, which is obtained from C_{t-1} and the input at the previous time step. If the forget gate f_t is 0, then the state at the previous moment is completely cleared, so that input data will be considered only with this time step. The input gate i_t determines whether to receive input at this time. The final output gate o_t determines whether to output the cell state. Hence, the overfitting is avoided by using RST in training data and selected the important features, which is used to improve the performance of RNN. The experiments and their validated results are discussed in next sections.

4. Results and discussion

In this section, the validation of proposed RST-RNN method and their experimental results are discussed with various existing techniques. Five biomedical datasets such as Pima Indians diabetes, Wisconsin breast cancer, heart disease, thyroid datasets and Parkinson datasets are collected from UCI machine learning repository [20] for identifying the performance of proposed RST-RNN method. Table 1 shows the details of dataset with ID, number of features and classes.

The missing values are present only in HD and BC dataset, the missing categorical attributes are replaced by using the mode of the attributes and the missing continuous data are replaced by mean of the attributes. During calculation, the numerical difficulties are addressed by scaling the data into the range of [-1,1] before constructing the proposed RST-RNN model. Hence, the feature values in the smaller numerical ranges are not dominated by those values present in the greater numerical ranges. In the following subsection, the evaluation of parameter settings with setup and the experimental validated results of RST-RNN method against various existing techniques are explained.

4.1 Experimental setup and parameter settings

The computer with 2.2 GHz of Intel Core i5, RAM of 8GB, where the RST-RNN method is developed using the programming language of Python 3.7.3 version. The performance of RST-RNN method is validated by conducting several experiments on UCI dataset using various metrics namely Area Under Curve (AUC), accuracy, F-measure, specificity (precision) and sensitivity (recall).

The proportion of positive samples that are correctly classified as positive by using sensitivity rate i.e. true positive rate. In contrast with this, the negative samples are correctly classified as negative by using specificity measure i.e. true negative rate. Accuracy can be calculated using the Eq. (11), and the Eq. (12) is used to evaluate the single combined metric, which is defined as F-measure. Among the number of labeled positive class samples, precision is used to identify the number of accurately labeled samples, which is shown in Eq. (13). On the contrary, according to the positive class, recall is used to predict the number of accurate positive class labeled samples, which can be divided by the total number of samples.

Table 1. Dataset description

Datasets with ID	No. of Classes	No. of instances	No. of features	Missing Values	No. of Samples for Training	No. of Samples for Testing
Heart Disease - HD	2	303	13	Yes	193	110
Breast Cancer - BC	2	699	9	Yes	499	200
Diabetes - PID	2	768	8	No	576	192
Parkinson - Pks	2	195	22	No	130	65
Thyroid - Thd	3	215	5	No	110	105

Table 2. Comparative analysis of proposed RST-RNN method

Methods	Accuracy(%)				
	HD	BC	PID	Pks	Thd
SVM+K-means	90.57	96.27	76.50	96.27	74.15
NB	97.36	95.64	76.48	96.24	76.48
RBF	96.77	95.57	76.47	95.45	77.19
J48	93.41	96.24	76.26	95.57	78.03
BPA-NB	97	95.12	78.50	96.12	78.29
Proposed RST-RNN	**98.57**	**98.19**	**81.47**	**98.15**	**83.46**

The mathematical expression for recall is given in Eq. (14).

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (11)$$

$$F-Measure = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (12)$$

$$Specificity = \frac{TN}{TN+FP} \quad (13)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (14)$$

Where, TP is true positive, TN is true negative, FP is false positive and FN is false negative.

4.2 Performance of proposed method by means of accuracy and AUC

In this section, the validation of RST-RNN method is analyzed against various existing techniques such as BPA-NB [13], SVM with K-means [19], hybrid approach such as Radial Basis Function as RBF, NB and J48. The existing BPA-NB conducted the experiments only on heart disease dataset.

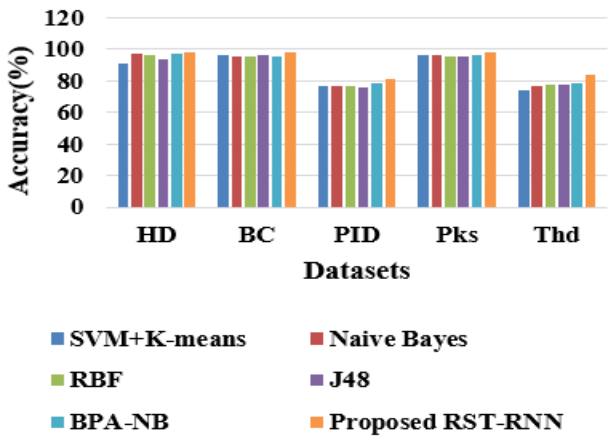

Figure. 2 Performance of RST-RNN method in terms of accuracy

Table 3. AUC performance of proposed method

Methods	AUC (%)				
	HD	BC	PID	Pks	Thd
SVM	Not Available	96.87	72.34	96.87	79.65
BPA-NB	87.05	98.45	85.40	96.13	84.41
Proposed Method	**89.58**	**99.12**	**87.46**	**97.67**	**89.74**

Therefore, to validate the RST-RNN method on various datasets, this algorithm implements the BPA-NB for other datasets and experiments are conducted. Table 2 shows the experimental results of RST-RNN method and graphical representation is given in Fig. 2.

From the Fig. 2, it is clearly stated that the RST-RNN method achieved better performance in all five datasets for accuracy parameters while comparing with existing techniques. For instance, the SVM with K-means techniques achieved nearly 77% accuracy for PID and Thd datasets, where the RST-RNN method achieved nearly 84% accuracy for those two datasets. The existing technique BPA-NB achieved nearly 97% accuracy for HD, BC and Pks datasets, where the RST-RNN method attained higher accuracy i.e. nearly 98.5% accuracy for the same datasets. The better performance of RST-RNN method is due to learning rate in RNN technique, which is used in proposed RST-RNN method. Table 3 provides the experimental results of proposed RST-RNN method against existing techniques: SVM and BPA-NB in terms of AUC for all five datasets. Fig. 3 illustrated the graphical representation of AUC performance.

From the Table 3 and Fig. 3, the experimental results showed that the RST-RNN method achieved better performance than popular existing techniques on all the different datasets. While comparing with other datasets, PID gained less AUC for all the three techniques.

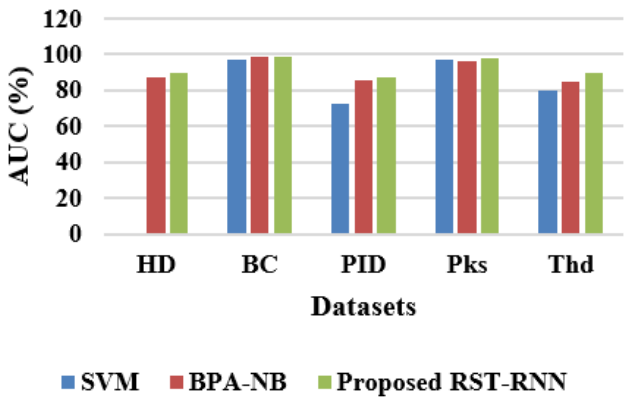

Figure. 3 AUC performance for proposed method

Table 4. Performance of proposed method against existing techniques

Methods	Sensitivity (%)					Specificity (%)				
	HD	BC	PID	Pks	Thd	HD	BC	PID	Pks	Thd
SVM	78.61	96.86	58.07	96.86	75.47	82.58	96.89	89.62	96.86	77.35
NB	97.40	96.24	59.59	96.42	79.13	97.90	96.59	88.80	96.59	78.26
RBF	97.07	96.27	64.13	94.57	79.87	96.23	94.32	88.16	94.32	80.19
J48	93.43	96.62	73.84	96.62	82.47	90.37	95.45	88.61	95.45	81.27
BPA-NB	96.51	97.50	80.49	95.14	85.68	84.16	98.25	78.46	95.47	87.12
Proposed Method	**98.24**	**99.53**	**94.51**	**98.47**	**90.47**	**98.64**	**99.74**	**96.39**	**98.50**	**89.95**

For instance, SVM achieved 72.34% AUC, BPA-NB achieved 85.40% AUC and RST-RNN method achieved 87.46% AUC. However, these techniques achieved higher AUC values for BC and Pks dataset.

4.3 Performance of proposed technique in terms of specificity and sensitivity

In this section, the parameters like specificity and sensitivity of RST-RNN method are compared with existing techniques such as SVM, RBF, NB, J48, and BPA-NB. The experimental results are tabulated in Table 4, in which the best values are make it as bold. Fig. 4 and 5 shows the graphical representation of sensitivity and specificity of RST-RNN method with several existing techniques. The sensitivity for all the datasets is experimented and the results are illustrated in Fig. 4.

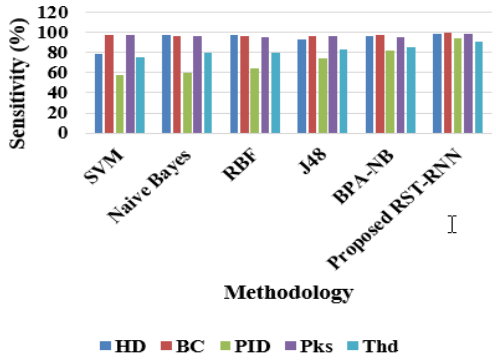

Figure. 4 Sensitivity of proposed method

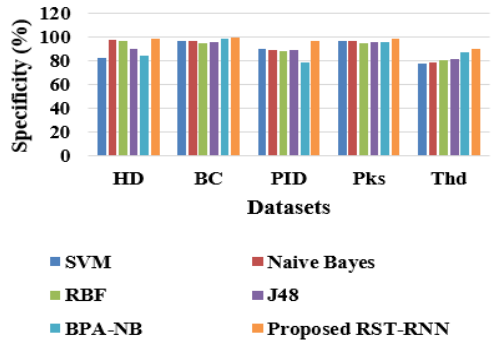

Figure. 5 Specificity of proposed method

It is clearly stated that the proposed RST-RNN method achieved higher performance than the existing techniques for all the datasets. The HD, BC and Pks achieved nearly 99% sensitivity for the RST-RNN method, whereas the existing techniques achieved nearly 96% sensitivity for RBF, J48, NB and BPA-NB techniques. When compared with other techniques, SVM provides poor performance on all other datasets except BC and Pks datasets. Fig. 5 shows the performance of specificity of RST-RNN method.

When compared with sensitivity of PID datasets, the specificity values have increased for the same dataset, which is illustrated in Table 4. But, the specificity values for Thd dataset provides low performance than other datasets for all techniques including RST-RNN method. For instance, the RST-RNN method achieved 89.95% specificity for Thd dataset and 99.74% specificity for BC dataset. When compared with SVM technique, the RST-RNN method improved the 7% specificity values for PID dataset. In the following sub-section, the performance of RST-RNN method in terms of F-measure are described.

4.4 Performance of proposed method by means of F-measure

The experiments are conducted on all dataset to validate the performance of RST-RNN method in terms of F-measure, which are shown in Table 5. The graphical representation for F-Measure of RST-RNN method are compared with BPA-NB, SVM and RBF is described in Fig. 6.

Table 5. Comparative analysis of proposed method

Methods	F-Measure (%)				
	HD	BC	PID	Pks	Thd
SVM	93.05	92.94	91.83	95.84	88.14
RBF	83.52	98.00	78.15	96.18	80.35
BPA-NB	91.25	95.12	90.81	94.70	87.34
Proposed Method	95.47	99.07	93.62	97.21	90.27

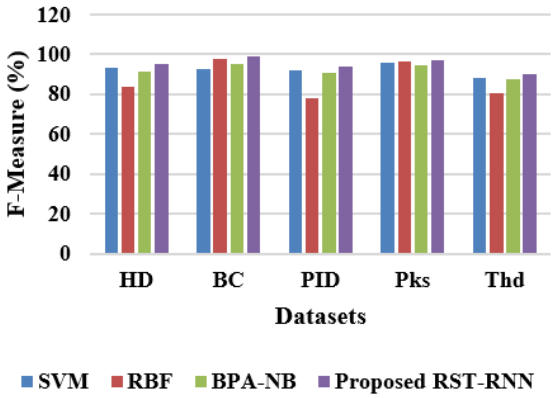

Figure. 6 Analysis of proposed method in terms of F-measure

The experimental analysis on F-Measure proved that the RST-RNN method achieved a higher F-measure than other existing methods for all five datasets. The RST-RNN method obtained 93.62% for PID dataset, whereas the RBF, BPA-NB and SVM achieved 78.15%, 90.81% and 91.83% F-Measure. In BC dataset, the RST-RNN method achieved 99.07% F-measure, where RBF achieved 98%. The analysis of the RST-RNN method in classification of five datasets shows that the RST-RNN method is highly efficient when compared to other existing methods such as SVM, RBF and BPA-NB. This shows that the RST-RNN method avoids the over-fitting of training data and selects the effective features using RST, which can be applicable for effective classification performance.

5. Conclusion

Nowadays, Big Data analytics plays a vital role in predicting diseases and tailoring of treatment for a particular disease. Big Data provides a 360-degree view of patients' data to perform analytics for better prediction outcomes. Prediction of healthcare increases the accuracy of diagnosis and helps to preventive medicine and public health. Predictive analytics with big data allow researchers to develop prediction models for accurate results over a large number of disease cases. However, the prediction using traditional methods are limited and time-consuming due to more number of features. In this research work, to address the issues of existing techniques, an RST-RNN method is developed to predict the diseases. The most important features are selected by using the RST technique and the classification for various diseases are carried out by RNN method. The experiments are conducted on five major UCI datasets in terms of several parameters to validate the effectiveness of RST-RNN against existing techniques. The proposed RST-RNN method achieved 98.57% accuracy, 89.58% AUC, 98.24% sensitivity, 98.64% specificity and 95.47% f-measure for HD dataset. The existing technique BPA-NB achieved accuracy of 97%, AUC of 87.05%, sensitivity of 96.51%, specificity of 84.16% and f-measure of 91.25% for the same HD dataset. The existing techniques didn't concentrate on the most relevant features; they process all the features for the classification. But, the proposed method designed an RSN method specifically to select the relevant features for effective classification. However, the performance of the proposed RST-RNN method achieved low performance in the Thd dataset, while comparing other datasets. In future work, this method is further improved by using effective optimization techniques to achieve higher performance in Thd UCI dataset.

References

[1] G. Manogaran and D. Lopez, "Health data analytics using scalable logistic regression with stochastic gradient descent", *International Journal of Advanced Intelligence Paradigms*, Vol.10, No.1-2, pp.118-132, 2018.

[2] A.N. Richter and T.M. Khoshgoftaar, "Efficient learning from big data for cancer risk modeling: A case study with melanoma", *Computers in biology and medicine*, Vol.110, pp.29-39, 2019.

[3] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities", *IEEE Access*, Vol.5, pp.8869-8879, 2017.

[4] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms", *Mobile Information Systems*, 2018.

[5] H. Hu, Y. Wen, T. S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial", *IEEE access*, Vol.2, pp. 652-687, 2014.

[6] A. Ed-daoudy, and K. Maalmi, "A new Internet of Things architecture for real-time prediction of various diseases using machine learning on big data environment", *Journal of Big Data*, Vol.6, No.104, 2019.

[7] M. Tarawneh and O. Embarak, "Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques", In: *Proc. of International Conference on Emerging Internetworking, Data & Web Technologies.* Springer, Cham, 2019.

[8] H. G. Schnack, "Improving individual predictions: machine learning approaches for

detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases)", *Schizophrenia research*, 2017.

[9] A. Ed-Daoudy and K. Maalmi, "Real-time machine learning for early detection of heart disease using big data approach", In: *Proc. Of International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS)*, 2019.

[10] H. D. Masethe and M. A. Masethe, "Prediction of heart disease using classification algorithms", In: *Proc. of the world Congress on Engineering and computer Science*, Vol.2, 2014.

[11] M. Kumari and S. Godara, "Comparative study of data mining classification methods in cardiovascular disease prediction", *International Journal of Computer Science and Technology*, Vol.2, No.2, pp.304-308, 2011.

[12] G. Purusothaman and P. Krishnakumari, "A survey of data mining techniques on risk prediction: Heart disease", *Indian Journal of Science and Technology*, Vol.8, No.12, pp.1, 2015.

[13] R. Venkatesh, C. Balasubramanian, and M. Kaliappan, "Development of Big Data Predictive Analytics Model for Disease Prediction using Machine Learning Technique", *Journal of Medical Systems*, Vol.43, No.8, pp.272, 2019.

[14] M. Nilashi, O. Ibrahim, H. Ahmadi, L. Shahmoradi, and M. Farahmand, "A hybrid intelligent system for the prediction of Parkinson's Disease progression using machine learning techniques", *Biocybernetics and Biomedical Engineering*, Vol.38, No.1, pp.1-15, 2018.

[15] A. Di Noia, A. Martino, P. Montanari, and A. Rizzi, "Supervised machine learning techniques and genetic optimization for occupational diseases risk prediction", *Soft Computing*, pp.1-14, 2019.

[16] L. R. Nair, D. Sujala Shetty, and D. Siddhanth Shetty, "Applying spark based machine learning model on streaming big data for health status prediction", *Computers & Electrical Engineering*, Vol.65, pp.393-399, 2018.

[17] T. Chen, J. Xu, H. Ying, X. Chen, R. Feng, X. Fang, and J. Wu, "Prediction of Extubation Failure for Intensive Care Unit Patients Using Light Gradient Boosting Machine", *IEEE Access*, Vol.7, pp.150960-150968, 2019.

[18] F. Van Wyk, A. Khojandi, and R. Kamaleswaran, "Improving Prediction Performance Using Hierarchical Analysis of Real-Time Data: A Sepsis Case Study", *IEEE Journal of Biomedical and Health Informatics*, Vol.23, No.3, pp.978-986, 2019.

[19] N. Kausar, A. Abdullah, B. B. Samir, S. Palaniappan, B. S. AlGhamdi, and N. Dey, "Ensemble clustering algorithm with supervised classification of clinical data for early diagnosis of coronary artery disease", *Journal of Medical Imaging and Health Informatics*, Vol.6, No.1, pp.78-87, 2016.

[20] A. Asuncion and D. Newman, "UCI machine learning repository", 2007.