JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

NEURAL NETWORK IN BIG DATA ANALYTICS FOR PREDICTION OF DISEASE EMPLOYING ROUGH SET THEORY

[1]Anusya.S, Amsaveni.M

[1]Assistant Professor,Assistant Professor
Department of Computer Science,
A.V.P College of Arts and Science, Tirupur,India.

Abstract : In today's world, early health prognosis is critical in preventing death due to treatment delays induced by forecast errors. Currently, academics are focusing on Big data, which is utilized to forecast future health conditions and provides an effective strategy to resolve initial forecasting difficulties. Several studies are being conducted on prediction analysis and modeling approaches to improve choice. Advanced analytics has significant potential for predicting future health status due to health factors and delivering the best results. Yet, owing to unclear or missing information in the database, information categorization is amongst the most difficult jobs. By eliminating unnecessary characteristics from the collected data, attribute selection approaches play an essential part in the classification. The Rough Set Theory (RST) approach is utilized in this study to choose more important characteristics that aids in the categorization of medical data and illness identification. For illness forecasting, the chosen characteristics are fed into the Neural Network (NN) method. The suggested approach is also known as RST-NN, and it involves conducting tests on the UCI machine and deep learning library data in terms of effectiveness and f-measure. For the cardiovascular disease data, the RST-NN approach obtained 98 percent accuracy, whereas the current Support Vector Machine (SVM) approach scored 91 percent efficiency, and Naive Bayes (NB) approach scored 97 percent accuracy.

IndexTerms **- Big data, Decision making, Feature extraction, Rough set theory, neural network**

I. INTRODUCTION

In several fields, such as health, science, and community, digital technology has become increasingly vital. Big data refers used to describe a significant volume of data that was collected and created from a variety of sources, including stream equipment, mobile applications, and, most notably, health. Due to the obvious lack of sufficient current technological devices, storing, displaying, and extracting research from various large data kinds becomes a difficulty. Identifying effective ways to get meaningful information on the various sorts of consumers is one of the most critical technical problems of data analytics [1].

Multiple kinds of health sources of data are now getting gathered both in medical and non-clinical settings, with the digital version of a personal health record is now the most significant data for health informatics. As just a result, there are 3 major obstacles to building a distributed information plan to deal with huge data: Firstly, because of the diverse and massive quantity of observations, it's indeed necessary to obtain data from dispersed places [2]. Secondly, with diverse and huge data, storage is the most significant issue. The issue is one of big data analytics, namely extracting massive datasets in real-time or near-real-time for modeling, forecasting, and optimization [3]. Because existing database management systems are inefficient in dealing with the diverse types of information or real-time [4] these issues necessitate a modern processing model.

The patients can die or become disabled as a result of an incorrect diagnosis. Medical experts can use the Health Predictions Framework to help them forecast a specific disease [5]. The massive quantity of data that can be acquired using electronic devices (by the client in the clinic) may be used in conjunction with big data to detect and forecast illnesses [6]. Because classification methods can analyze huge volumes of information, techniques are frequently utilized in healthcare. NB, SVM, Closest Neighbour, logistic regression, Fuzzification, Fuzzy dependent computational model, Machine learning algorithm, and genetic programming are some of the most commonly utilized approaches in healthcare [7]. For complicated metrics, machine learning with a categorization may be used effectively in medical uses. For illness prediction, modern categorization approaches enable more sophisticated and efficient forecasting methodologies [8]. The relevant characteristics are picked in this study utilizing the RST approach, which is utilized to improve the categorization method's effectiveness. The NN technique is used to categorize the information using the essential characteristics of the data set. Overfitting within the dataset is also decreased when employing the RST approach, and the suggested RST-NN approach is validated on several UCI datasets for illness prediction.

II. LITERATURE REVIEW

In [9] developed BPA-NB, a Big Data Predictive Analytic Tool for predicting heart disease based on the NB method. To analyze the data, this system employed a probabilistic categorization based on Bayes' theory. BPA-NB utilized the network algorithm to filter out extraneous information and generating an accurate forecast. Through using the MapReduce technique with the Apache Spark platform, the complexity was decreased. The studies were conducted performed on the UCI dataset to compare BPA-NB to current methods in terms of computation, CPU utilization, and reliability. The Incremental Linear Regression Model (ILRM) for the Universal Parkinson's Disease Grading Test was designed in [10]. ILRM was tasked with predicting Engine and Total-UPDRS. A self-organizing map was utilized to combine the data, and a non-linear recurrent fractional least-square was employed to reduce complexity. Numerous empirical studies were performed using a real-world PD format compatible with UCI to assess the ILRM technique.

[11] used pattern classification and computational methods, such as SVM and KNN, to forecast workplace illness risks. To use the k-means method, the model equation was judged to be a collection of coherent labeled groupings. The ideal hyperparameter and optimum ad-hoc heterogeneity measure weighting for categorizations were discovered using evolutionary algorithms, which enhanced the effectiveness of such systems. [12] used a hybrid strategy for health information categorization, combining SVM with the K-means segmentation method. The Based Feature Extraction technique was used to decrease the characteristic dimensions. Then, to distinguish between normal and pathological patients, the associated characteristics and measurements were successfully modified. In terms of effectiveness, accuracy, memory, and f-measure, the trials have been performed on the UCI datasets. The created research lowered the classification performance with higher recognition duration when the unknown pattern of comparable behavior was added into the sampled communities.

III. PROPOSED METHODOLOGY

Occur as a result of numerous relationships here between the history of patients and the sickness, illness forecasting using health information is a vital task. Illness prediction has several benefits, including diagnostics and a lower mortality rate. There was a significant amount of health records, which needed to be evaluated properly. The RST and NN are used to forecast illness in medical data in this study. The dependence here between attributes is discovered by evaluating the characteristics' features with RST, which is also used to eliminate unnecessary characteristics. The NN was given the RST-generated decision function as inputs. For illness forecasting, the NN evaluates the characteristics of the data using a decision function. This chapter would go over the ins and outs of how RST and NN function in depth. Figure 3.1 shows a schematic diagram of RST and NN in illness forecasting.

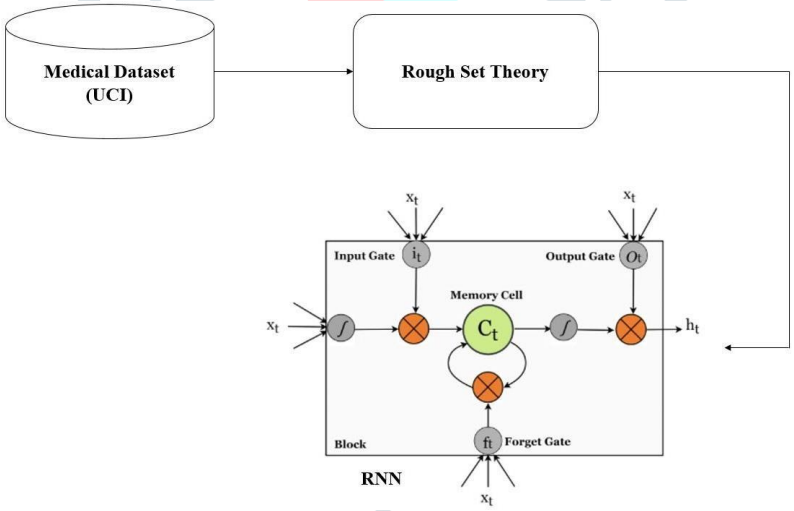

Figure 3.1: Framework of RST and NN

3.1 Rough Set Theory

Let I = (Ú, X) denote an information management system, with Ú denoting a non - empty set collection of finite objects known as the universe of discourse and A denoting a non - empty set of characteristics. Each characteristic with the letter A has a system of values (Vo) connected with it. There is a related inductive argument IND (Q) for a subgroup of characteristics X, which is known as an unnormalized relationship. Equation (3.1) may be used to define the relationship IND (Q):

$$(Q) = \{(x, y) \in Ú\ 2 | \forall o \in Q, O(x) = O(y)\} \quad (3.1)$$

If (x, y) IND(Q), x and y are indistinguishable by characteristics from P. [x]P denotes the similarity matrix of the P-indiscernibility relationship. The RST's theoretical foundation is the unnormalized relationship. In RST, the bottom and top approximation are two fundamental procedures. X denotes a subgroup. By building the P-lower approximation denoted as QX, X may be an approximation using just information provided inside Q. QX is the collection among all items that can be categorically categorized as members of X based on the characteristic set Q. The Q- higher approximations of X, indicated as QX, that may be categorized as components of X using the feature value P. The expression is shown as Equation (3.2) & (3.3).

$$QX = \{X|[X]Q \subseteq QX\} \quad (3.2)$$

$$QX = \{X|[X]Q\ Ú\ X \neq \varphi\} \quad (3.3)$$

Here QX represents Q-lower estimate and QX represents Q-higher approximation. The RST identifies features that are dependent on characteristics and eliminates those that are unnecessary. The RST's characteristics are fed into the NN's classification method as inputs.

3.2 Neural Network

The challenge of representing changes that occur in time-series data is well-suited to NNs. Natural language, speech synthesis, and writing pattern recognition all employ them. The time change vector sequence Yt1, Yt, Yt+1,.. is fed into the NN. As the series progresses, the input Yt and the preceding hidden layer At1 impact the hidden state St at the same time. To properly explain the NN procedure, are using the equations (3.4) and (3.5) below:

$$At = (U.\ Yt + W.\ At-1) \quad (3.4)$$

$$Ot = (V.\ At) \quad (3.5)$$

Wherever, At denotes the pattern's recollection at period t, i.e. the value of the hidden units, as determined by Equation (4). W is the preceding moment's result, which is utilized as the weighted intake at this time, and U is the input's sampling value. Equation (5) is utilized to determine the output as Ot with V describing the output sample weight. Both f and g are activation functions, with f being a tanh, ReLU, or sigmoid transfer function. A softmax function is usually denoted by the letter g.

The slope computed by hidden state backpropagation can disappear or expand as the NN architecture develops. Gradient cropping can deal with gradients eruptions, but it can't deal with gradient disappearing. As a result, NN cannot readily represent the dependency among text components across long distances in a speech model's text series. The abovementioned issues can be solved by using a long short-term memory (LSTM). The condition of a unit has been at the heart of an LSTM. It has three different types of gate structures: inputs, outputs, and forgetting. The necessary Equation (3.6-3.10) formulae continues to follow:

$$ft = \sigma(Wf.\ [ht-1, xt] + bf) \quad (3.6)$$

$$it = \sigma(Wi.\ [ht-1, xt] + bi) \quad (3.7)$$

$$ot = \sigma(Wo.\ [ht-1, xt] + bo) \quad (3.8)$$

$$Ct = ft \times Ct-1 + it \times \tanh(Wc.\ [ht-1, xt] + bC) \quad (3.9)$$

$$ht = ot \times \tanh(Ct) \quad (3.10)$$

There are three multiplication gateways in Equation (3.6-3.8): the forget gate, ft; the input value, it; as well as the outlet port, ot. In Equation (3.6 -3. 8) the input is [xt, ht-1], but the variables are varied. The sigmoid activation function is represented as. In Equation (3.9), Ct is the convolution layer, which would be calculated from Ct1 and the preceding time step's input. When the recall gates ft is set to 0, the prior state is entirely wiped, but just this time step's input is evaluated. The input gate decides not whether input will be received at this moment. The output data gate ot decides whether or not the cell should be transmitted. As a result, by utilizing RST in training examples and selecting the relevant features, overfitting is prevented, and the efficiency of NN is improved. In the following parts, we'll go over the tests and their verified findings.

IV. RESULTS AND DISCUSSION

The implementation of the proposed RST-NN approach and its experimental findings are detailed in this part in comparison to different current methodologies. Five medical data, including Pima Indian diabetes, Wisconsin breast cancer, cardiovascular disease, endocrine sets of data, and Parkinson sets of data, were gathered from the Machine learning repository [13] to assess the suggested RST-NN techniques efficiency. The dataset's ID, quantity, and categories are listed in Table 4.1.

Just the HD and BC datasets have incomplete data; lacking category characteristics are substituted with the median of the attributes, and lacking maintained constant is substituted with the mean of the attributes. Before creating the suggested RST-NN model, the numerical challenges are handled by scale all information into the range of [-1,1]. As a result, obtained features in the lower numeric range are not overshadowed by those in the larger numeric range. The assessment of parameters with installation as well as the experimentally confirmed outcomes of the RST-NN approach versus different current methods is explained in the next part.

4.1 Experimental setup and parameter settings

The RST-NN technique was created using Python 3.7.3 programming on a machine with a 2.2 GHz Intel Core i5 processor and 8GB of RAM. The accuracy, F-measure, selectivity (accuracy), and sensitivities of the RST-NN technique are confirmed by running multiple tests on the Collected information utilizing multiple indicators such as Area Under The curve (AUC), accuracy, F-measure, and sensibility.

True positive rate is the proportion of positive samples that are correctly identified as positives using the sensitivities ratio. Negative specimens, on either hand, are accurately categorized as negatives using sensitivity measures, i.e. true negative frequency. Equation (4.1) is used to compute precision, and Equation (4.2) is being used to assess the single integrated metric, which would be defined as F-measure. Accuracy is utilized to determine the number of properly labeled samples among the number of labeled affirmative class instances, as indicated in Equation (4.3). Recall, on either hand, can be used to estimate the number of the correct

positive category labeled data, which may then be split by the entire sample, according to the true positive. The equation gives the statistical model for recall (4.4).

$$Accuracy = \frac{TPS+TNS}{TPS+FPS+FNS+TNS} \quad (4.1)$$

$$F-Measure = \frac{2\times Precision\times Recall}{Precision+Recall} \quad (4.2)$$

$$Specificity = \frac{TNS}{TNS+FPS} \quad (4.3)$$

$$Sensitivity = \frac{TPS}{TPS+FNS} \quad (4.4)$$

While TP denotes genuine positivity, TN denotes true negativity, FP denotes false positivity, and FN denotes false negativity.

The reliability and AUC of the suggested technique were used to identify areas of improvement.

The previous BPA-NB experimented only on datasets related to heart disease. As a result, the approach implements the BPA-NB for additional datasets, and tests are undertaken to verify the RST-NN technique on various data. Figure 2 shows the graphical depiction of the experimental findings of the RST-NN technique. When compared to current approaches, the RST-NN method achieved superior efficiency in all multiple datasets on precision metrics, as shown in Figure 3.2. Again for PID and Thd data, for example, the SVM using K-means methods obtained almost 77 recognition rate, while the RST-NN technique produced 85 recognition rate. For the HD, BC, and Pks data, the current approach BPA-NB obtained roughly 98 recognition rates, but the RST-NN method achieved nearly 98.5 recognition rates for the same datasets. The learning rate of the NN approach, which is employed in the suggested RST-NN technique, is responsible for the enhanced quality of the RST-NN technique.

Table 4.1: Dataset for the study

Entity	Class No	Instances No	Features No	Training Samples	Test Samples
Heart Disease - HRD	3	333.3	14.3	212.3	121
Breast Cancer - BIC	3	768.9	9.9	548.9	220
Diabetes - PIID	3	844.8	8.8	633.6	211.2
Parkinson - PkS	3	214.5	24.2	143	71.5
Thyroid - ThD	4	236.5	5.5	121	115.5

The pictorial depiction of AUC effectiveness is shown in Figure 4.1. The experimental outcomes in Figure 4.1 indicated that the RST-NN number of advantages over traditional prominent current methods on all data. PID has worse AUC for any three methods when compared to certain other data.

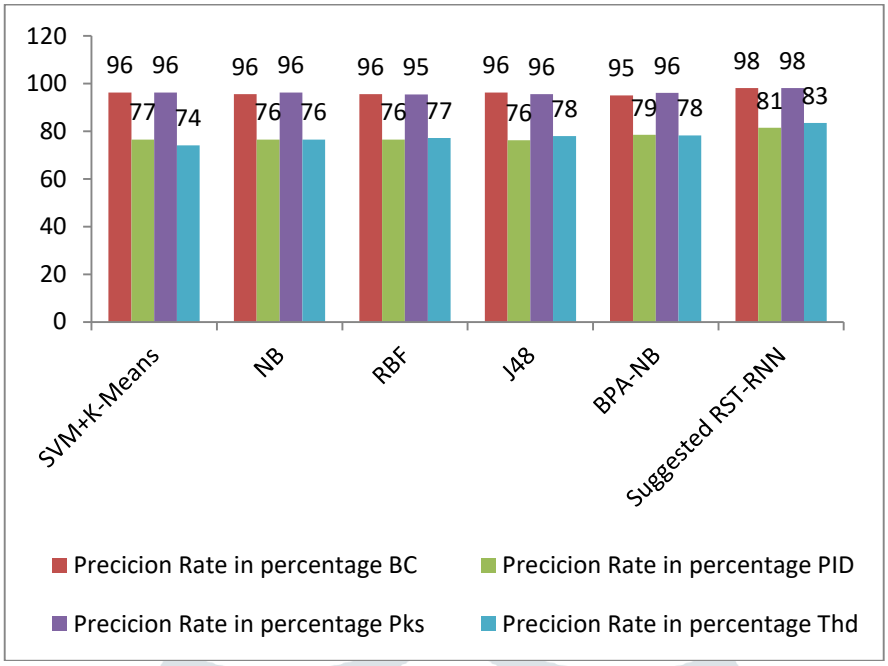

Figure 4.1: Accuracy Rate of RST-NN

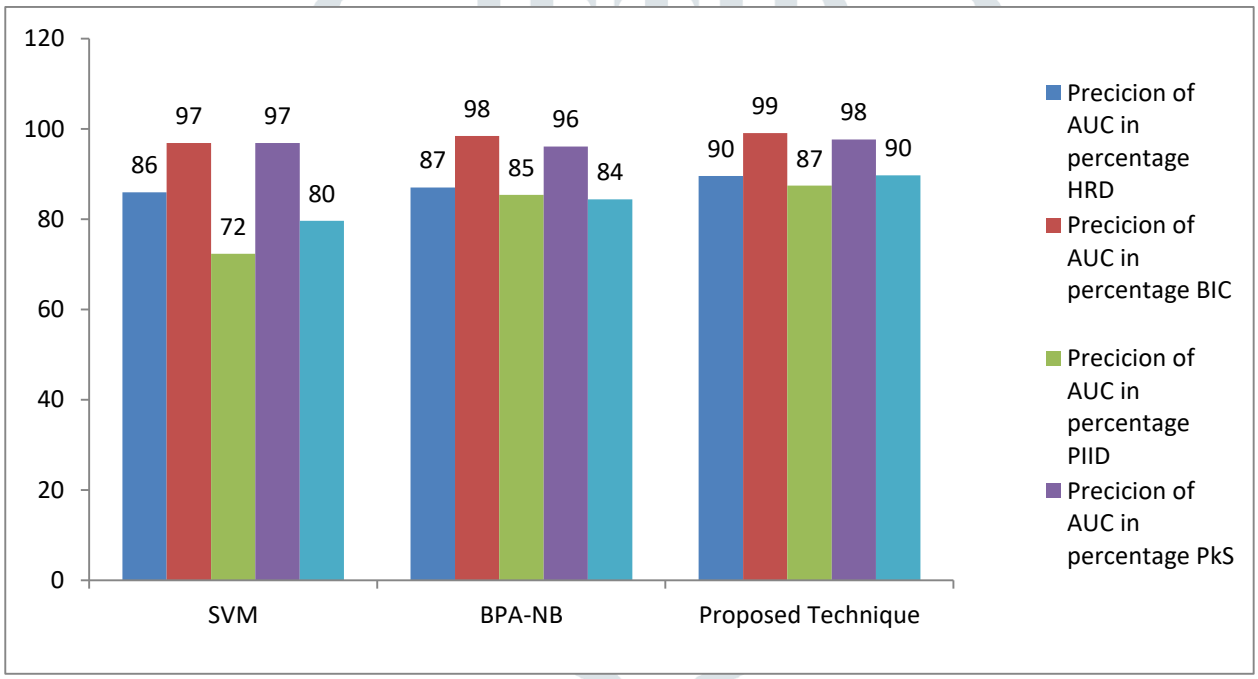

Figure 4.2: AUC efficacy of the suggested technique

4.2 The suggested method's performance as measured by the F-measure

Experimental tests were carried out on all datasets to verify the RST-NN technique's results in terms of F-measure, as seen in Table 4.2. Figure 4.2 provides a graphical depiction of the F-Measure of an RST-NN technique when comparing to BPA-NB, SVM, and RBF.

Table 4.2 shows a comparison of the proposed technique

Methods	**F-Measure (%)**				
	HRD	**BIC**	**PIID**	**PkS**	**ThD**
SVM	93	93	92	96	88
RBF	84	98	78	96	80
BPA-NB	91	95	91	95	87
ProposedMethod	95	99	94	97	90

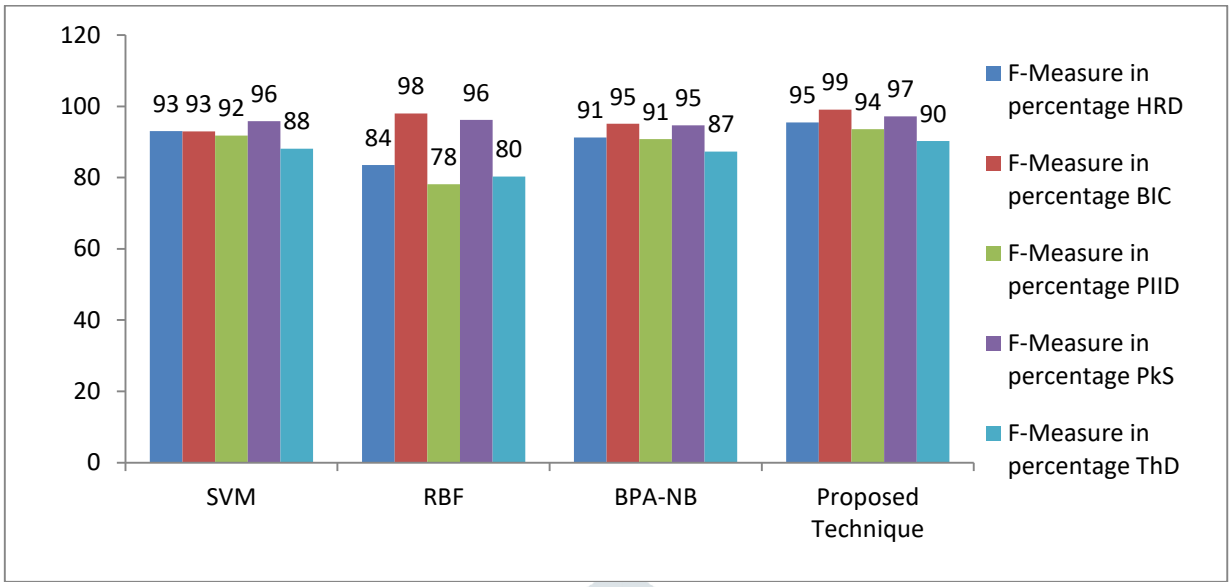

Figure 4.3: F-measure evaluation of the suggested technique.

For all five data, the experiment conducted on F-Measure revealed that the RST-NN approach obtained a higher F-measure than some other current techniques. Again for PID data, the RST-NN technique received 93.62 percent, while the RBF, BPA-NB, and SVM methods received 78.15 percent, 90.81 percent, and 91.83 percent F- Measure, respectively. The RST-NN technique got 99.07 percent F-measure on the BC data, whereas RBF got 98 percent. As contrasted to certain other current techniques such as SVM, RBF, and BPA- NB, the RST-NN method is very effective in classifying data sets, according to the research. This demonstrates that the RST-NN technique prevents over-fitting of training information and utilizes RST to pick the most useful features, which may be used to improve classification results.

V. CONCLUSIONS

Data Analytics is critical for forecasting illnesses and customizing treatment for a specific condition. Big Data gives a 360º picture of a patient's information, allowing for improved analytics and forecast results. Health forecasting improves diagnosis accuracy and aids preventative healthcare and population health. Researchers may use prediction big data analysis data to make prediction models that provide reliable findings over a large number of illness cases. Traditional approaches, on the other hand, are restricted and time-consuming given a large number of characteristics. To overcome the drawbacks of previous approaches, an RST-NN algorithm for illness prediction is proposed in this study. The RST approach is used to pick the most relevant characteristics, and the NN technique is used to classify the various illnesses. The tests compare the efficacy of RST-NN to existing methods on five large UCI datasets in terms of numerous metrics. For the HD dataset, the suggested RST-NN approach has 98 percent accuracy, 90 percent AUC, 98 percent sensitivity, 99 percent specificity, and 95 percent f-measure. For identical HD data, the current approach BPA-NB obtained 97 percent accuracy, 87 percent AUC, 97 percent sensitivities, 84 percent specificity, and 91 percent f-measure. The current approaches did not focus on the most important traits; instead, they processed all of them just for categorization. The proposed technique, on the other hand, developed an RSN method to pick the important characteristics for successful categorization. Yet, as compared to certain other datasets, the efficiency of the suggested RST-NN approach was poor in the Thd data.

REFERENCES

[1] Vijayakumar, D.R., Arjunan, K.P., Sivasakthi, M. and Lakshmanan, K., 2019. DIABETES PREDICTION BY MACHINE LEARNING OVER BIG DATA FROM HEALTHCARE COMMUNITIES. IRJET Apr.
[2] Haq, A.U., Li, J.P., Memon, M.H., Nazir, S. and Sun, R., 2018. A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Information Systems, 2018.
[3] Hu, H., Wen, Y., Chua, T.S. and Li, X., 2014. Toward scalable systems for big data analytics: A technology tutorial. IEEE access, 2, pp.652-687.
[4] Ed-daoudy, A. and Maalmi, K., 2019. A new Internet of Things architecture for real-time prediction of various diseases using machine learning on big data environment. Journal of Big Data, 6(1), pp.1-25.
[5] Tarawneh, M. and Embarak, O., 2019, February. Hybrid approach for heart disease prediction using data mining techniques. In International Conference on Emerging Internetworking, Data & Web Technologies (pp. 447-454). Springer, Cham.
[6] Schnack, H.G., 2019. Improving individual predictions: machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases). Schizophrenia research, 214, pp.34-42.
[7] Kumari, M. and Godara, S., 2011. Comparative study of data mining classification methods in cardiovascular disease prediction 1.

[8] Purusothaman, G. and Krishnakumari, P., 2015. A survey of data mining techniques on risk prediction: Heart disease. Indian Journal of Science and Technology, 8(12), p.1.
[9] Venkatesh, R., Balasubramanian, C. and Kaliappan, M., 2019. Development of big data predictive analytics model for disease prediction using machine learning technique. Journal of medical systems, 43(8), pp.1-8.
[10] Nilashi, M., Ibrahim, O., Ahmadi, H., Shahmoradi, L. and Farahmand, M., 2018. A hybrid intelligent system for the prediction of Parkinson's Disease progression using machine learning techniques. Biocybernetics and Biomedical Engineering, 38(1), pp.1-15.
[11] Di Noia, A., Martino, A., Montanari, P. and Rizzi, A., 2020. Supervised machine learning techniques and genetic optimization for occupational diseases risk prediction. Soft Computing, 24(6), pp.4393-4406.
[12] Kausar, N., Abdullah, A., Samir, B.B., Palaniappan, S., AlGhamdi, B.S. and Dey, N., 2016. Ensemble clustering algorithm with supervised classification of clinical data for early diagnosis of coronary artery disease. Journal of Medical Imaging and Health Informatics, 6(1), pp.78-87.
[13] Asuncion, A. and Newman, D., 2007. UCI machine learning repository.