International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1S4, December 2019

Dice Similarity Based Gaussian Deep Recurrent Neural Learning for Classification and Prediction with Big Data Analytics

S Arun Kumar, M Venkatesulu

Abstract: ***Big data analytics is a process of gathering large volume of data and organizing the present and past events to predict future events. Analyzing such a huge volume of data is not a simple task. Therefore, processing large data is a challenging one to predict an accurate event. The conventional techniques handling the large volume of data but the accurate prediction was not obtained since it failed to progressively learn the higher level features from raw inputs. An efficient Dice similarity based Gaussian Deep Recurrent Neural Learning Classifier (DS-GDRNLC) model is developed to enhance the prediction performance in terms of prediction time, prediction accuracy with big data. Initially, DS-GDRNLC model gathers huge volume of data from the big dataset (DS). After that, the gathered data are trained with several layers such as input layer, two hidden layers and output layer. The numbers of data are given to the input layer for performing the classification. Then the proposed DS-GDRNLC model uses two hidden layers to repeatedly learn the input data using a regression function. The regression function uses the dice similarity coefficient to find the relationship between the data and the predicted class. Then the analyzed results at the hidden layers are fed into the output layer. The Gaussian activation function is used at the output layer to verify the similarity value and mean of class. If the similarity value is closer to the mean of class, then the data are classified into that specific class. In this way, all the input data are accurately classified into the different classes resulting improves the Prediction Accuracy (PA). Finally, the training error rate is calculated for each classification results for obtaining the higher PA. This process repeated until the minimum error is obtained. Experimental evaluation is performed with big DS using different metrics such as PA, precision, recall, F-measure and Prediction Time (PT). The observed results confirm that the DS-GDRNLC model efficiently increases the PA, precision, recall as well as F-measure and minimizes the PT than the state-of-the-art methods.***

Keywords: ***Predictive analytics, big data, Gaussian Deep Recurrent Neural Learning Classifier, regression, dice similarity coefficient, Gaussian activation function***

I. INTRODUCTION

With the increase of big data, the prediction is a part of statistics that extracting more relevant information from the large volume of data and predict the future and behavior patterns. In recent days, big data is the fastest and more broadly used in several application fields such as, healthcare, weather, agriculture, business, marketing, banking and so on. Recently, several research works have been developed for predictive analytics with big data. However, it faces several challenges for accurate prediction.

The support vector trained multilayer neural network (SVM- trained MNN) was developed in [1] for classifying the big data with the help of map-reduce function. The method minimizes the PT but the accurate classification was not achieved. In [2], a CNN-based multimodal disease risk prediction (CNN-MDRP) algorithm was presented with big data. The designed algorithm increased the PA but it has higher time to predict the disease.

A statistical assessment model of the healthcare information system was presented in [3] for predictive analysis with big data. Though the accuracy and F-measure were improved, the time consumption was not minimized using statistical assessment model. Artificial Intelligence (AI) methods were developed in [4] to predict disease with Big Data. But the methods failed to perform accurate disease prediction. For predicting malaria disease, an artificial neural network (ANN) was introduced in [5]. But the error rate of the prediction was not minimized. A deep learning paradigm was developed in [6] for increasing the health prediction using big data. But the repeated data learning was not performed to achieve higher accuracy.

In [7], an ensemble multi-label classification technique was designed to identify the disease risk at an earlier stage. But the designed technique failed to minimize the error. A spark based machine learning technique was developed in [8] for remote health status prediction. The designed technique failed to predict a variety of diseases.

A machine-learning-based diagnosis scheme was introduced in [9] to predict heart disease with minimum computation time. But the performance of disease prediction was not improved while considering the large volume of data. A decision tree classification technique was developed in [10] to improve the performance of heart disease prediction. But, the false positive rate (FPR) was not minimized.

Abovementioned issues are overcome by presenting a novel DS-GDRNLC model.The major contribution of the proposed DS-GDRNLC model is summarized as follows

- To improve the PA, DS-GDRNLC model is introduced by repeatedly learning the input data in the hidden layers. In the hidden layers, the input data are analyzed using the regression function. To find the relationship between dependent variable (i.e. data) and independent data (i.e. predicted class), the regression function uses the dice similarity coefficient.

Manuscript published on 30 December 2019.
***** Correspondence Author (s)
S Arun Kumar*, Department of Computer Science and Enginnering, Kalasalingam Academy of Research and Education, Krishnankoil, India. Email: arunsphd@gmail.com
M Venkatesulu, Department of Information Technology, Kalasalingam Academy of Research and Education, Krishnankoil, India.
Email: m.venkatesulu@klu.ac.in

Based on the similarity values, the data are classified at the output layer. This helps to improve the precision, recall and F-measure.

- To minimize the PT, the Gaussian activation function is used at the output layer for analyzing the similarity value and mean value of the predictive class. The similarity value which is closer to the mean is classified into the particular class.

The rest of the paper is structured into five different sections. Section 2 reveals the related works. In Section 3, the proposed DS-GDRNLC model is explained with the architecture diagram. In Section 4, experimental evaluation is performed with big data. The experimental results are discussed with different parameters in Section 5. Section 6 provides the conclusion of the paper.

II. RELATED WORKS

Machine learning methods with rule-based classification were introduced in [11] for forecasting a variety of liver diseases. But the accurate prediction was not achieved with minimum time. A deep neural network (DNN) and long-short-term memory (LSTM) learning models were developed in [12] for predicting various infectious diseases. But the prediction performance was not increased.

A big data analytics-enabled transformation method was introduced in [13] and applied to the healthcare organization for predicting the disease. The performance of accuracy and time complexity remained unsolved. An efficient and privacy-preserving disease risk prediction (EPDP) system was developed in [14] for disease risk prediction. The system failed to improve the accuracy of diagnosis results at an earlier stage.

A probabilistic data acquisition technique was designed in [15] to predict the future health status of the patients. Though the designed technique improves the PA, the PT was not minimized. The three machine learning classification techniques were developed in [16] to predict diabetes at an early stage. The algorithms failed to predict or diagnose other diseases.

A distributed association rule-based classification method was developed in [17] based on MapReduce for big data analytics. The method improves the accuracy but the error rate was not minimized. A predictive analysis algorithm using Hadoop/Map Reduce was developed in [18] to forecast different types of diabetes with high accuracy. But the PT was not minimized.

Predictive analytics with big data in emergency care was developed [19] to predict the disease. But the performance of the disease prediction remained unsolved. A recurrent convolutional neural network (RCNN) was developed in [20] to evaluate the disease risk. The disease risk prediction time was not minimized.

Major issues in big data predictive analysis are less prediction accuracy, high prediction time and error rate. These existing issues are overcome by presenting a novel DS-GDRNLC model. The brief explanation of DS-GDRNLC model is presented in the next section.

III. METHODOLOGY

The DS-GDRNLC Model is developed for predicting future events with a large volume of data. The management and analysis of big data are the most significant emerging needs as the huge volume increasing the complexity of data being created or collected. Therefore, deep recurrent neural learning is introduced to effectively perform predictive analytics. The "deep learning" in the proposed DS-GDRNLC model refers to use the number of hidden layers for deeply analyzing the input data through which the data is transformed to obtain better outcomes.

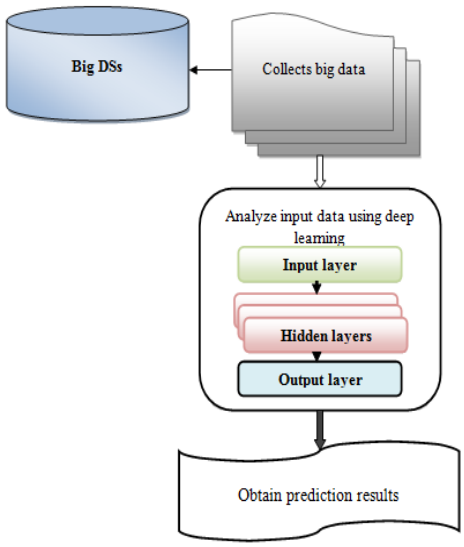

Figure 1 architecture of DS-GDRNLC model

More precisely, deep learning is the sequence of transformations from the input to output. The normal neural learning uses the one input layer, one hidden and one output layer for processing the input big data. But the deep learning includes one input layer, two or more hidden layers and one output layer. The different hidden layers are used to deeply learn the input data and provide accurate prediction results at the output layer by using regression function. A function of the independent variables called the regression function is to be estimated. Regression fuction focus is on the relationship between a dependent variable (i.e. Class) and one or more independent variables (i.e. data') Regression Fuction helps one understand how the typical value of the dependent variable (i.e. Class) changes when any one of the independent variables is varied, while the other independent variables are held fixed. The processing diagram of the proposed DS-GDRNLC model is shown in figure 1.

Figure 1 depicts the architecture of the proposed DS-GDRNLC model to obtain better prediction results with minimum time complexity. Initially, the big DS is considered for predictive analytics. The number of data is taken from the big DS. Using deep recurrent neural learning, the data analysis is performed with the input data. The input data is fed into the input layer of the deep architecture. Then the incoming data are analyzed in the hidden layers. Finally, the results are obtained at the output layer. The structural diagram of the deep recurrent neural leering technique is shown in figure 2.

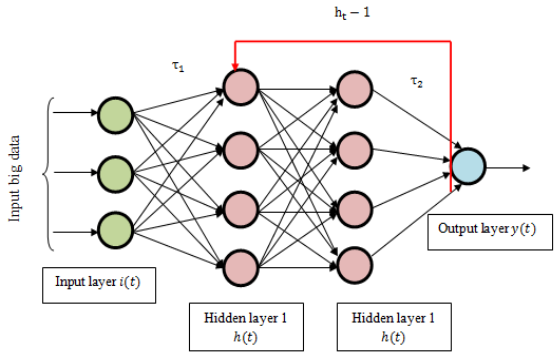

Figure 2 structure of the deep recurrent neural learning

Figure 2 illustrates the structure of the deep recurrent neural learning with the one hidden layer, two hidden layers and output layer. The numbers of big data are given to the input $i(t)$ layer $d_1, d_2, d_3, \ldots . d_n$ at time't' . The deep recurrent neural learning includes the neurons-like the nodes into different layers and interconnected directly into the next successive layers with adjustable weightsτ_1,τ_2. The nodes in the one layer are fully connected to the consequent layers and repetitively perform the deep learning of input data hence the name is called as Recurrent Deep neural network. The inputs are transformed into the hidden layers. The adjustable weights between the input and hidden layer is represented as τ_1 as shown in figure 2. The input layer of the deep learning is expressed as,

$$i(t) = \sum_{i=1}^{n} d_i(t) * \tau_1 \quad (1)$$

In (1),$i(t)$ represents the input, $d_i(t)$ denotes a big data at a time't', τ_1 denotes an adjustable weights between the input and hidden layer. In the proposed deep learning, there are two hidden layers are used for predicting the future results with big data. The regression is used in hidden layers to analysis the input data. The similarity between dependent variable (i.e. input data) and one or more independent variables are evaluated by the regression analysis. The variable whose values to be predicted is known as dependent variable and independent variable is number of input data. The relationship is measured based on the similarity function. The dice similarity is employed for finding relationship between dependent variable and independent variable which is computed as follows,

$$\beta = 2 * \frac{d_i \cap c_i}{|d_i| + |c_i|} \quad (2)$$

In (2), β represent the regression coefficient, d_i represents the independent variable (i.e. data), c_i denotes a predicted dependent variable (i.e. class). $|d_i|\ and\ |c_i|$ represents the cardinalities of the two sets (i.e. a number of a elements in each set). The dice similarity is equals to the ratio of the mutual independence between the two variables and the sum of the number of elements in each set. The dice similarity coefficient (β) provides the similarity value between 0 and 1. Based on the regression analysis, the output of the hidden layer is represented as '$h(t)$'.

$$h(t) = \tau_1 * i(t) + \tau_h * h(t-1) \quad (3)$$

In (3) $h(t)$denotes a hidden layer output at a time't'. Here, '$h(t-1)$' represents a unit time delay output from hidden layer 2 to hidden layer 1 and 'τ_h' denotes a weights of the hidden layer, τ_1 represents a weight between input and hidden layers, $i(t)$ represents the input. The output of the second hidden layer is feedback into the input of the first hidden layer and repeatedly learning the input data and given to the output layer.

Then the similarity values are transformed into the output layer where the similarity values are verified and predict the results. The hidden layer output is transformed into the output layer.

$$y(t) = \gamma_f(\tau_2 * h(t)) \quad (4)$$

In (4),$y(t)$ represents the output of the deep recurrent neural learning, τ_2 represents the adjustable weights between the hidden layer and output layer, $h(t)$ denotes an output of the hidden layer. γ_f denotes a activation. The proposed DS-GDRNLC model uses the Gaussian activation function to classify the data into a particular class uses the mean and deviation.

$$\gamma_f = e^{\left(-\frac{(\beta-\mu)^2}{2\sigma^2}\right)} \quad (5)$$

In (5), β denotes a similarity value, μ denotes a mean of the class, 'σ' represents the deviation from the mean. If the similarity value is closer to the mean, the activation function returns '1'. It means that the data are classified into the particular class. Otherwise, the similarity is deviated from the mean of a particular class and the activation function returns '0'. In this way, all the data are correctly classified into the particular class. After the classification, the error rate is computed for obtaining the higher PA. The error rate is calculated based on the difference between the squared difference between the actual and predicted results. It is mathematically formulated as follows,

$$E = \{y_a(t) - y(t)\}^2 \quad (6)$$

In (6),E denotes an error, $y_a(t)$ denotes an actual output, $y(t)$ represents the predicted result. Based on the error rate, the weights of the connection between the layers are adjusted and the process is repeated until it reaches the minimum training error.

$$T_e = \arg\min E \quad (7)$$

In (7),$T_e$represents the minimum training error, $\arg min$ denotes an argument of the minimum function, E represents the error. This process minimizes the incorrect prediction of the future events and enhances the prediction accuracy. The algorithmic process of the DS-GDRNLC model is described as follows,

Input : Big dataset, Number of data $d_1, d_2, d_3 \ldots d_n$.
Output: Improve prediction accuracy
Begin
1. Give d_i to input layer $i\ (t)$ with the weight τ_1
2. Transform input data into hidden layer $h(t)$
3. **For** each data
4. Compute the similarity between d_i and c_i
5. Transform $h(t)$ into $y(t)$
6. **If** $(y_f = 1)$ **then**
7. Classified the data into the particular class
8. **Else**
9. Classified the data into the different class
10. End if
11. Calculate training error E
12. **Update the** weights of the connection between the layers $\Delta\tau_1, \Delta\tau_2$
13. The process is iterated until find $argmin\ E$
14. **End for**
End

Algorithm 1 Dice similarity based Gaussian deep recurrent neural learning

Algorithm 1 describes the DS-GDRNLC for predictive analytics with higher accuracy and minimum time. The numbers of input data are given to the input layer. The input data are transferred into the hidden layer for analyzing the input data and their predicted classes. For each data, the similarity between the input data and their respective predicted classes are calculated in the hidden layers. The data are repeatedly learned in the two hidden layers. The similarity values are fed into the output layer. In the output layer, the similarity values are verified with the class mean value. The proposed classifier uses the Gaussian activation function. The activation function effectively verifies the similarity value and mean of the particular class. The similarity which is close to the mean is classified into a particular class. Finally, the training error is calculated for each result and the processes are continued until the minimum error is attained. Thus, deep learning enhances PA and lessens the FPR.

IV. EXPERIMENTAL SETTINGS

To validate the DS-GDRNLC technique, the comparison is made with existing methods SVM-trained MNN [1], CNN-MDRP [2], RCNN (recurrent convolution neural network) [3] are implemented in Java Language to predict the diseases at an earlier stage. Big data (i.e. patient's information) are gathered from Cardiovascular Disease DS (https://www.kaggle.com/sulianova/cardiovascular-disease-DS) and Pima Indians Diabetes Database (https://www.kaggle.com/uciml/pima-indians-diabetes-database).

Cardiovascular Disease DS comprises 70,000 records of patient's data with 13 attributes. This DS is used to identify the conditions that involve blocked blood vessels that direct to a heart attack, chest pain (angina) or stroke. The attributes are id, age, gender, height, weight, ap_hi (systolic blood pressure), ap_lo (diastolic blood pressure), cholesterol, glucose, smoke, alco, active and cardio.

The Pima Indians Diabetes Database is used to predict the patient affected by diabetes or not, based on certain analytical measurements information's. In this DS, all the patients here are females with 21 years old of Pima Indian heritage. The DS comprises the 9 attributes such as pregnancies, Glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age, and outcome.

The experimental evaluation is done with the above said medical DSs for predicting the two diseases with the patient information's. The different parameters such as PA, precision, recall, F-measure and PT are evaluated with number of patient files.

V. RESULTS AND DISCUSSION

In this section, Results and discussion of DS-GDRNLC and existing methods namely SVM-trained multilayer neural network (MNN) [1], CNN-MDRP [2], RCNN [3] are described. Various metrics namely PA, precision, recall, F-measure and PT with number of files are presented to validate the DS-GDRNLC technique in healthcare industry.

A. Performance analysis of prediction accuracy

PA is measured as ratio of number of patient files are correctly predicted as disease or not to total number of input files taken for the experimental. PA is formalized as follows,

$$prediction\ accuracy = \frac{Number\ of\ files\ are\ correctly\ predicted\ as\ disease\ or\ not}{n} * 100 \quad (8)$$

In (8),n denotes a number of the patient files. The PA is measured in terms of percentage (%).

PA of three different methods is computed with number of patient files. While considering 20 patient files, the DS-GDRNLC technique correctly predicts 17 patient files as disease or not and PA is 85%. The PA of the other two methods is 75% and 70% respectively. Followed by, the nine runs are carried out with a number of patients' files. The PA of the proposed DS-GDRNLC technique is improved when compared to existing SVM-trained MNN [1] , CNN-MDRP [2] and RCNN [3] respectively. The ten different results are shown in figure 3.

Table 1. Tabulation for prediction accuracy

Number of patient files	Prediction accuracy (%)			
	DS-GDRNLC	SVM-trained MNN	CNN-MDRP	RCNN
20	85	75	70	50
40	90	80	73	55
60	92	83	72	57
80	91	80	70	61
100	94	82	76	63
120	92	83	75	65
140	91	86	75	68
160	94	81	75	70

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

180	91	82	74	73
200	95	84	80	75

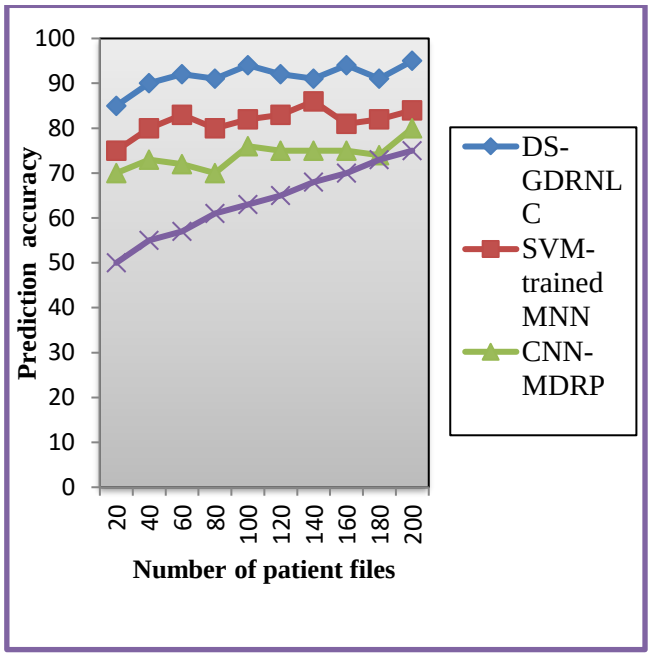

Figure 3 simulation results of number of patient files vs prediction accuracy

Figure 3 portrays results of PA versus number of patient files which is ranges from 20 to 200. Patient information's is gathered from big DS. The numbers of patient files is taken as input in 'x' direction and PA is obtained at 'y' direction. The above graph clearly shows that the DS-GDRNLC technique improves PA than the existing classification techniques. This improvement is achieved by applying a deep recurrent neural network to effectively learn the patient information's such as systolic blood pressure, diastolic blood pressure, cholesterol, glucose, insulin, BMI and so on. This patient information's are fed into the input layer. Then the input data are transformed into the hidden layers and repetitively learned using regression analysis and transferred to the output layer. The output layer uses the Gaussian activation function to correctly classify the patient files as cardiovascular disease and Diabetes disease. In this way, the diseases are correctly predicted with higher accuracy.

The ten different results of PA are compared with the existing classification techniques. Therefore, DS-GDRNLC technique enhances the PA by 12%, 24% and 45% as compared to existing [1], [2]and [3] respectively.

B. Performance analysis of precision

Precision is measured as number of patient files is accurately predicted as disease to the total number of input files taken for experimental evaluation. Precision is formulated as follows,

$$Precision = \frac{TP}{Tp+FP} * 100 \quad (9)$$

In (9),TP denotes a true positive, FP represents the false positive. True positive is numbers of files are correctly predicted as disease and FPR is numbers of files are incorrectly predicted as disease. Precision is calculated in percentage (%).

Table2. Tabulation form precision

Number of patient files	Precision (%)			
	DS-GDRNLC	SVM-trained MNN	CNN-MDRP	RCNN
20	87	77	67	60
40	94	86	74	63
60	96	88	71	65
80	95	87	78	67
100	97	88	83	70
120	95	89	84	73
140	93	85	81	75
160	96	88	80	77
180	95	87	79	79
200	97	89	85	82

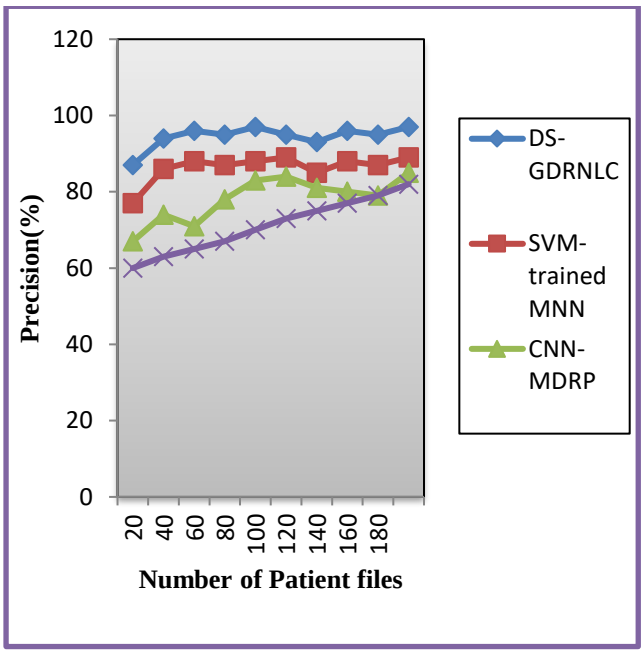

Figure4. Simulation results of number of patient file vs precision

Table 2 portrays the results of precision with number of patient files. Totally ten different results are described in table 1. For each iteration, the various precision results are obtained. The results show that the DS-GDRNLC technique improves precision when compared to the SVM-trained MNN [1], CNN-MDRP [2] and RCNN [3] respectively. The proposed DS-GDRNLC technique efficiently predicts cardiovascular disease and Diabetes disease at an earlier stage. By utilizing the deep learning classifier, the disease prediction is performed. Through the regression analysis, the incoming patient files are analyzed. The regression function analyzes the patient data and predicted class using the Dice similarity coefficient. Based on the similarity value, the Gaussian function classifies the disease types in an accurate manner. The error rate is calculated after the classification to reduce the FPR and enhances the true positive rate.

The ten different precision results of DS-GDRNLC technique is compared with the precision of the conventional classification techniques. The comparison results clearly show that the precision is significantly increased by 9%, 21% and 33% than the existing SVM-trained MNN [1], CNN-MDRP [2] and RCNN [3] respectively.

C. Performance analysis of Recall

Recall is measured as ratio of the true positive to the sum of true positive and false negative results from the input files. The recall is formulated as follows,

$$Recall = \frac{TP}{Tp+FN} \quad (10)$$

In (10),TP denotes a true positive, FN represents the false negative. True positive is the numbers of files are correctly predicted and false negative is the numbers of files are incorrectly predicted as normal. Recall also measured in terms of percentage (%).

The above table values are described that the evaluation results of recall using three different classification techniques namely DS-GDRNLC technique, SVM-trained MNN [1], CNN-MDRP [2] and RCNN [3]. For the experiential purposes, the number of patient's files taken as input in the range of 20 to 200 to compute the recall. The reported result confirms that the recall using proposed DS-GDRNLC technique is higher as compared to other classification techniques. The DS-GDRNLC technique performs deep learning with the patient data and predicts the disease affected by the patient at an earlier stage based on the similarity coefficient value. This in turn helps for increasing the patient files are correctly predicted at the output layer and minimizing the true negative rate. Therefore, the DS-GDRNLC technique improves the recall then the conventional technique. Because the conventional techniques failed to deeply analyze the patient data for accurate prediction.

Table 3. Tabulation for recall

Number of patient files	Recall (%)			
	DS-GDRNLC	SVM-trained MNN	CNN-MDRP	RCNN
20	93	83	73	60
40	94	86	83	62
60	95	92	84	65
80	96	91	83	67
100	97	90	87	70
120	96	91	86	72
140	98	87	85	75
160	97	90	84	77
180	97	88	87	79
200	97	92	88	81

Let us consider the 20 patient files as input, true positive rate of DS-GDRNLC technique is 13 and the false negative rate is 1, then the percentage of recall is 93%. But the true positive rate and the false negative rate of SVM-trained MNN [1], CNN-MDRP [2] and RCNN [3] are 10, 2 and 8, 3 respectively. Then the percentages of a recall are 83% and 73% respectively. Similarly, nine different results are obtained with a number of patients files. The ten different results of DS-GDRNLC technique is compared to the existing techniques. The results confirm that the recall is considerably increased by 8%, 15% and 36% then the state-of-the-art methods.

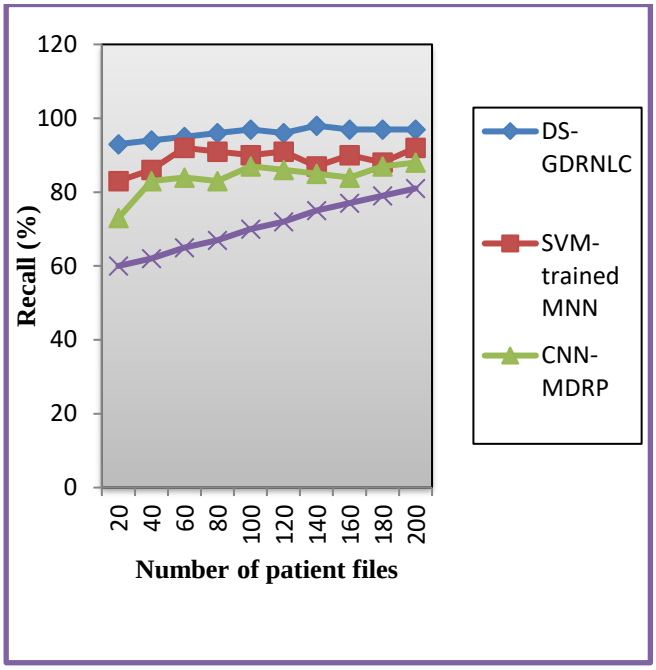

Figure 5. Simulation results of number of patients vs recall

D. Performance analysis of F-measure

F-measure is a measure of an evaluation accuracy of the classification algorithm and it is the average of the precision and recall. The F- measure is mathematically calculated as follows,

$$F - measure = 2 * \frac{Pr*Rc}{Pr+Rc} * 100 \quad (11)$$

From equation (11), Pr represents the precision, Rc denotes a recall. F- measure is measured in percentage (%).

Table4. Tabulation for F-Measure

Number of patient files	F- Measure (%)			
	DS-GDRNLC	SVM-trained MNN	CNN-MDRP	RCNN
20	90	80	70	63
40	94	86	78	65
60	95	90	85	67
80	95	89	80	70
100	97	89	85	72
120	95	90	85	75
140	95	86	83	78
160	96	90	82	80
180	96	87	83	82
200	97	90	86	84

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Table 4 demonstrates the results of F-measure versus number of patient files ranges from of 20-200. For predicting the disease at an earlier stage, patient information's are gathered from DS and maintained in the file format. From the table value, it is proved that the proposed DS-GDRNLC model provides better F-measure results as compared to existing classification techniques. This is because of DS-GDRNLC model increases the precision as well as recall in the disease prediction analysis with big patient data. The F-measure is a measure of a test's accuracy of the DS-GDRNLC model which is obtained both precisions as well as recall.

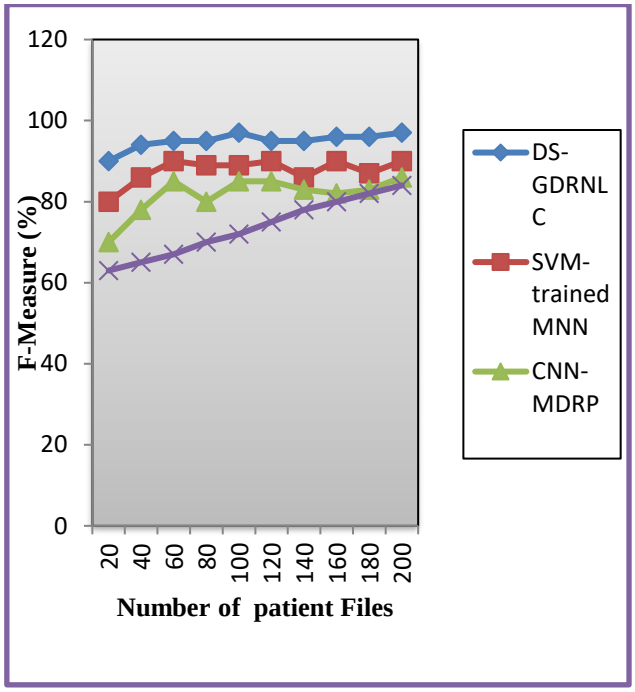

Figure 6 Simulation results of number of patient files vs F-Measure

While considering the input count of the patient file is 20, the precision and recall of DS-GDRNLC model are 87% and 93%. Then the percentage of F-measure is 90%. Similarly, the F-measure of the other two techniques is 80%and 70% respectively. Similarly, the patient files counts are increased for each run and the results are obtained. The performance results show that the DS-GDRNLC model outperforms well and the F-measure is improved by 8%, 17% and 30% than the SVM-trained MNN [1], CNN-MDRP [2] and RCNN [3] respectively. and 70% respectively. Similarly, the patient files counts are increased for each run and the results are obtained. The performance results show that the DS-GDRNLC model outperforms well and the F-measure is improved by 8%, 17% and 30% than the SVM-trained MNN [1], CNN-MDRP [2] and RCNN [3] respectively.

E. Performance analysis of prediction time

Prediction time is measured as amount of time taken to predict the disease type based on patient current information (i.e. file). PT is mathematically calculated as follows,

$$prediction\ time = n * T\ (predictingthedisease\ ) \quad (12)$$

In (12),n represents a number of files, T denotes a time for predicting the disease. PT is measured in terms of milliseconds (ms).

Table 5. Tabulation for Precision Time

Number of patient files	Precision time(%)			
	DS-GDRNLC	SVM-trained MNN	CNN-MDRP	RCNN
20	17	20	24	26
40	22	24	28	29
60	25	29	31	31
80	30	34	37	38
100	34	38	42	43
120	37	43	47	49
140	40	45	49	52
160	44	48	51	54
180	47	50	54	58
200	50	54	58	60

Figure 5 portrays the results of PT of proposed and existing methods with number of patient files. The above graphical result shows that the PT is said to be minimized when compared to the existing methods. By increasing the number of patient files, with the time taken to predict the disease also increases. But comparatively, the PT is found to be lesser usingDS-GDRNLC model due to the application of DS-GDRNLC. This classifier analyzes the input patient data at the two hidden layers and prediction results are obtained at the output layer. In that layer, the Gaussian activation function is used for analyzing the similarity value and mean value of the predictive class. If the similarity value is more close to the mean value, then the patient file is classified into the specific class. The classification results are used for predicting the disease with minimum time. The results of the PT of DS-GDRNLC model is minimized by 11% compared to the SVM-trained MNN [1] and 19% compared to CNN-MDRP [2] and 28% compared to RCNN [3].

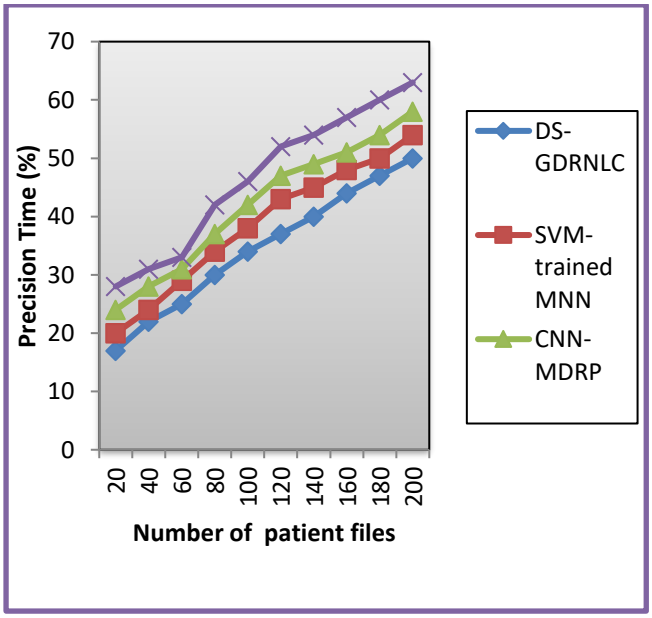

Figure 7 simulation results of number of patient files vs precision time

The above discussion of the various performance metrics clearly shows that the DS-GDRNLC model effectively improves the disease prediction with higher accuracy and minimum time.

VI. CONCLUSION

An efficient model called DS-GDRNLC is introduced to achieve higher PA and minimal time with the big data. The data are collected from the big DSs and given to the input layer of the recurrent deep neural learning classifier. Then the inputs are learned repetitively in two hidden layers using the regression analysis with dice similarity coefficient. The learned data are transferred to the output layer. In the output layer, the Gaussian activation function is used for classifying the data based on the similarity and mean value. Finally, the training error is calculated and the process is repeated until the minimum error is found. This helps to improve the accuracy and minimizes the FPR. Experimental evaluation is carried out using healthcare big DSs with the parameters such as PA, precision, recall, F-measure and PT. The observed result clearly shows that DS-GDRNLC model improves the disease PA, precision, recall, F-measure and minimizes the PT when compared to the state-of-the-art methods.

ACKNOWLEDGMENT

The first author is thankful to the management of Kalasalingam Academy of Research and Education for providing fellowship.

REFERENCES

1. Ahmad Ali AlZubi, “Big data analytic diabetics using map reduce and classification techniques”, The Journal of Supercomputing, Springer, April 2018, pp 1–10
2. Min Chen, Yixue Hao, Kai Hwang, Lu Wang and Lin Wang, “Disease Prediction by Machine Learning over Big Data from Healthcare Communities”, IEEE Access, Volume 5, April 2017, pp 8869 – 8879
3. C. B. Sivaparthipan, N. Karthikeyan and S. Karthik, “Designing statistical assessment healthcare information system for diabetics analysis using big data”, Multimedia Tools and Applications, Springer, November 2018,pp 1–14
4. Zoie S.Y.Wong, Jiaqi Zhou, Qingpeng Zhang, “Artificial Intelligence for infectious disease Big Data Analytics”, Infection, Disease & Health, Elsevier, Volume 24, Issue 1, 2019,pp 44-48
5. Thakur Santosh and Dharavath Ramesh, “Artificial neural network based prediction of malaria abundances using big data: A knowledge capturing approach”, Clinical Epidemiology and Global Health, Elsevier, Volume 7, 2019, pp 121–126
6. Hongye Zhong and Jitian Xiao, “Enhancing Health Risk Prediction with Deep Learning on Big Data and Revised Fusion Node Paradigm”, Scientific Programming, Hindawi, Volume 2017, June 2017, pp1-18
7. Runzhi Li,Wei Liu, Yusong Lin, Hongling Zhao, and Chaoyang Zhang, “An Ensemble Multilabel Classification for Disease Risk Prediction”, Journal of Healthcare Engineering, Hindawi, Volume 2017, June 2017, pp 1-10
8. Lekha R.Nair, Sujala D.Shetty, Siddhanth D.Shetty, “Applying spark based machine learning model on streaming big data for health status prediction”, Computers & Electrical Engineering, Elsevier, Volume 65, January 2018, pp 393-399
9. Amin Ul Haq , Jian Ping Li , Muhammad Hammad Memon , Shah Nazir ,and Ruinan Sun, “A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms”, Mobile Information Systems, Hindawi, Volume 2018, December 2018, pp 1-21
10. Jaymin Patel, Tejal Upadhyay, Dr. Samir Patel, “Heart Disease Prediction Using Machine learning and Data Mining Technique”, International Journal of Computer Science & Communication, Volume 7, Issue 1,2015, pp 129-137
11. Yugal Kumar and G. Sahoo, “Prediction of different types of liver diseases using rule based classification model”, Technology and Health Care, Volume 21, Issue 5, 2013, pp 417–432
12. Sangwon Chae, Sungjun Kwon and Donghyun Lee, “Predicting Infectious Disease Using Deep Learning and Big Data”, International Journal of Environmental Research and Public Health, Volume 15, 2018, pp 1-20
13. Yichuan Wang, LeeAnn Kung, William Yu Chung Wang, Casey G. Cegielski, “An integrated big data analytics-enabled transformation model: Application to health care”, Information & Management, Elsevier, Volume 55, Issue 1, 2018, pp 64-79
14. Xue Yang , Rongxing Lu , Jun Shao , Xiaohu Tang , Haomiao Yang, “An Efficient and Privacy-Preserving Disease Risk Prediction Scheme for E-Healthcare”, : IEEE Internet of Things Journal, Volume 6 , Issue 2 , 2019, pp 3284 – 329
15. Prasan Kumar Sahoo , Suvendu Kumar Mohapatra ,Shih-Lin Wu , “Analyzing Healthcare Big Data With Prediction for Future Health Condition”, IEEE Access, Volume 4, 2016, pp 9786 - 9799
16. Deepti Sisodia and Dilip Singh Sisodia, “Prediction of Diabetes using Classification Algorithm”, Procedia Computer Science, Elsevier, Volume 132, 2018, pp 1578–1585
17. Alessio Bechini, Francesco Marcelloni, Armando Segatori, “A MapReduce solution for associative classification of big data”, Information Sciences, Elsevier, Volume 332, 2016, pp 33-55
18. N.M. Saravanakumar, T.Eswari, P.Sampath, S.Lavanya, “Predictive Methodology for Diabetic Data Analysis in Big Data”, Procedia Computer Science, Elsevier, Volume 50, 2015, pp 203-208
19. Alexander T.Janke BS, Daniel L.Overbeek MD, Keith E.Kocher MD, MPH, Phillip D.Levy MD, MPH, “Exploring the Potential of Predictive Analytics and Big Data in Emergency Care”, Annals of Emergency Medicine, Elsevier, Volume 67, Issue 2, 2016, pp 227-236
20. Mohd Usama , Belal Ahmad , Jiafu Wan , M. Shamim Hossain , Mohammed F. Alhamid , M. Anwar Hossain, “Deep Feature Learning for Disease Risk Assessment Based on Convolutional Neural Network With Intra-Layer Recurrent Connection by Using Hospital Big Data”, IEEE Access , Volume 6, 2018, pp 67927 - 67939

AUTHORS PROFILE

S. Arunkumar was born in Srivilliputtur, Tamil Nadu, and India. He received the B.E. degree in Computer Science and Engineering from the Anna University, Tamil Nadu, and India. In 2010, the M.Tech. Degree in Software Engineering from Karunya University, Coimbatore, Tamil Nadu, India. He has held Assistant Professor Positions at Kalaivani College of Technology, Coimbatore, Tamil Nadu, and India. And the University of Petroleum and Energy Studies, Dehradun, Uttarakhand. He is currently a Research Scholar at Kalasalingam Academy of Research and Education, Krishnankovil, Srivilliputtur, Tamil Nadu, and India. His research interests include Big Data Analytics, Distributed Computing, and Machine Learning.

M.Venkatesulu received the postgraduate degree in Mathematics from Sri Venkateswara University, Tirupati, India, in 1975, and the Ph.d degree in mathematics from Indian Institute of Technology, Kanpur, India, in 1979. He worked as a faculty member at Shri Sathya Sai University, Prashanthinilayam, India Between 1983 and 2003. He also worked as a consultant for Satyam Computers, Hyderabad, India, for short period. He was Visiting Professor at University of Missouri, Kansas City, between August 2006 and May 2007. Currently he is working as a Senior Professor and Head of the Department of Information Technology at Kalasalingam Academy of Research and Education, Krishnankovil, Srivilliputtur, Tamil Nadu, and India. His area of interest includes differential equation, Image Processing, Cryptography, Bioinformatics, Big Data Analytics and Distributed Computing.