International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-3, February 2019

Performance Evaluation of MachineLearning Techniques in Diabetes Prediction

Raghavendra S, Santosh Kumar J, Raghavendra B. K.

Abstract*: Diabetes diagnosis is very important at preliminary stage rather than treatment. In today's world devices like sensors are used for detection of diabetes. Accurate classification techniques are required for automatic identification of diabetes disease. In regards to research diabetes prediction with minimal number of attributes (test parameters) is to be identified earlier research states about feature reduction but with less predictive accuracy. In this regards, this work exploits machine learning techniques(methodology) such as Logistic Regression (LR), Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF) and Neural Network (NN) with 10-fold Cross Validation (CV) for classification and prediction of diabetes with Feature Selection Methods (FSMs) using R platform. Above all models enable us to investigate the relationship between a categorical outcome and a set of explanatory variables. The experiment was conducted on PIMA Indian diabetes dataset selected from UCI machine learning repository. From the experimental results it is identified that for full set of diabetes dataset attributes, Classification Accuracy (CA) achieved was 84.25%whereas with reduced set attributes an accuracy of 85.24% is achieved using NN with 10-fold CV technique compared to others which will help in medical application to predict diabetes with minimal features.

Keywords – Logistic regression; Artificial neural network; Random forest; Support vector machine; Neural network with 10-fold.

I. INTRODUCTION

Recent improvements in the area of data mining and machine learning have empowered the research in biomedical field to improve the condition of general health care. In many parts of the world the tendency for maintaining long-lasting records consisting of medical data is becoming an accepted practice. In addition to this, the newer medical equipment's and the techniques used in diagnosis, produces composite and huge data. Therefore, to handle these ill-structured biomedical data, intelligent algorithms for data mining and machine learning are required in order to take logical reasoning from the saved raw data, which is considered as medical data mining. Within the medical data, the medical data mining searches for patterns and relationships which can provide useful information for appropriate medical diagnosis [1]. Data mining techniques are applied to different medical domains (health care databases or medical datasets) to improve the medical diagnosis. LR is one of the techniques used to evaluate a relationship among two variables, one is a non-independent variable consisting of categorical values and other is a non-dependent variable [2]. LR is used in studying datasets, where there are one or more non-dependent variables to check the results and the outcome can fail/pass, false/true, absent/present etc. [3].

ANN is one among the various fields of Artificial Intelligence. The human brain architecture is the main inspiration behind the development of the model. ANNs are successfully used in various disciplines such as environmental science, study of human mind, study of numbers, study of medicine, study of computers etc. ANNs are also being used in many business areas like accounts and audits, funding, managing and decision making, promotion and manufacture etc. ANNs have turned out to be a well-liked model and recently they are used to identify diseases and to forecast the patients' survival proportion [4].

In machine learning SVMs are the models used for supervised learning accompanying with other learning algorithms which can analyze data used for regression and classification. For any set of training examples given, each of them is marked as fitted to one or other group, an SVM training algorithm constructs a model that allocates new examples to a single category or the other, constructing it a non-probabilistic binary linear classifier[5].

To check for any invisible patterns inside the medical datasets, medical data mining is strongly recommended. In medical data mining, the actual tasks (challenges) are the classification and prediction of medical datasets. One of the techniques that are used for the classification and prediction is random forest [6]. Random forest is an ensemble learning method for regression, classification, and other jobs that functions by making an assembly of decision trees at training time and generating the class that is the classification or regression of the distinct trees. Random forest is a versatile algorithm used for classification suited for the analysis of large datasets. Random forest is popular because random forest classification models have high prediction accuracy and provides information on important variables for classification. Random forest provides two important aspects for data mining i.e. high prediction accuracy and information related to importance of attribute in classification [7].

In 10-fold CV the main step is to split dataset into k chunks and run neural network k times using a different chunk for testing each time. The other k-1 chunks are used for training. Now for that particular network architecture you have k different sets of results which, combined, have been tested across the entire data-set. These results can be averaged and errors can be calculated to produce an accurate range for the performance of the network.

Manuscript published on 28 February 2019.
***** Correspondence Author (s)
Dr. Raghavendra*, Associate Professor in the Department of Computer Science and Engineering at Christ Deemed To Be University, Bangalore.
Santosh Kumar, Associate Professor in the Department of Computer Science and Engineering at K.S.School of Engineering and Management, Bangalore.
Dr. Raghavendra B.K. Pursued P.hd From VTU Belgaum, Karnataka and Masters from VTU Belagavi and Bachelors from Bengaluru

II. LITERATURE SURVEY

From the existing literature we found many different methods are applied on PIMA Indian diabetes dataset. The methods proposed by different researcher send the classification accuracy achieved are explained below:

To exhibit the efficiency of the hybrid classifier based on evolutionary computation on diabetes dataset, a method based on hybrid classifier along with k-nearest neighbor was proposed. Based on the classification accuracy, it was clear that on over 50 runs the hybrid classifier achieved good accuracy of 80% [8]. Instead of using a traditional neuron which produces output for a given input in each iteration, a spiking neuron which gets activated after each T ms with an input is designed. The output can be changed into a particular firing rate furthermore it can perform the data classification depending on a firing rate created from input signal. For a set of cases belonging to one among k classes, every input is connected to input current and the spiking neuron gets excited after T ms, at last the firing amount is calculated for each case. Weights for the spiking neuron are optimized using a gravitational search algorithm. The capability of the projected method is compared with the identical spiking neuron implemented with particle swarm optimization (PSO), cuckoo search algorithm and differential evolutions. The model is implemented on diabetes dataset and the gravitational search algorithm achieved good accuracy of 76.61% [9]. For optimizing the parameter for SVM, an adjusted bat algorithm (ABA)is proposed. The experiments are conducted on the diabetes dataset. The experimental result was compared with the Grid-SVM and other approaches. Based on the result, ABA-SVM is considered as a better classifier than Grid-SVM and compared to other approaches like PSO-SVM, the ABA-SVM achieved better classification accuracy of 77.34% [10]. A model is proposed to handle the problems that can appear when learning from very small data that are already classified. The model depends on Logical Analysis of Data (LAD) and is provided with additional information obtained from the consideration of data statistically. So the new proposed model is called SLAD. The performance of SLAD is compared with LAD, SVM and label propagation algorithm. The experiment was conducted on diabetes dataset. From the results obtained, it was found that for both 5% training and 10% training SLAD achieved better accuracy of 72.87% compared to the other methods [11]. A method to use sequential variational inference and kalman filtering on diabetes dataset to predict the classification accuracy is proposed [12]. From the output of the method, it was clear that sequential variational inference achieved better accuracy of 80% compared to 76% achieved by kalman filtering. By combining the advantages of graph and combinatorial method, a clustering ensemble method was developed using Dempster-Shefer evidence theorem [13]. The model was implemented on diabetes dataset and from the experimental results it was identified that the proposed theorem achieved better accuracy of 69.27% compared to other methods. A method similar to principal component analysis was used to select the important attributes was developed [14]. These attributes are given as an input to the feed forward artificial neural network. The result achieved by the method is measured up with other methods of the feature selection like Tarr's, RUCK's, principal component analysis and t-test. The new model was applied on the diabetes dataset. Testing is done using 20% of data and remaining 80% is used for training. The proposed method achieved good accuracy of 75.22% with less number of attributes. A selective Bayesian classifier is proposed and is implemented on diabetes dataset using 5-fold cross validation sample [15]. The augmented Bayesian classifier is also implemented on the same dataset. The result of selective Bayesian classifier is compared with naïve bayes and augmented Bayesian classifier. From the result it was clear that selective Bayesian classifier gives better accuracy than the naïve bayes and augmented Bayesian classifier. In addition to this, the selective Bayesian classifier achieves better accuracy of 79.94% through lesser amounts of attributes thus by reducing the size of the dataset. For inductive concept learning an Evolutionary Concept Learner (ECL) was developed and three different selection mechanisms of ECL: US (US selection operation), WUS (Weighted US) and exponentially weighted US (EWUS) were implemented on diabetes dataset [16]. From the result it was found that the average accuracy achieved by EWUS was 77% and better than compared to US and WUS. A model that makes use of genetic algorithm to select important features is developed in parallel with mapreduce framework [17]. The selected features are produced to k-Nearest Neighbor classifier. The experiment is carried out on diabetes dataset. The accuracy of fitness is calculated using k-Nearest Neighbor. From the result it was seen that parallel genetic algorithm produces better accuracy of 80.51%. A powerful method is proposed for low dimensional classification and estimation of regression problems [18]. Classification difficulty may be considered as a difficulty of approximating the training set. A multi resolution framework is built based on approximations and organized in the form of a tree. This supports for efficient training. The model is experimented on diabetes dataset and achieves a good accuracy. An artificial immune recognition system which can notice the existence or nonexistence of disease is developed. The diabetes dataset is run on the machine on an average of 3 runs using 10-fold cross validation sample. The capability of the model is compared with the supplementary methods like IncNEt, Logdisc and Dipol92. The accuracy obtained by the proposed system was 74.1% and was better than the others [19]. A growing-pruning spiking neuron network consisting of 2 stage learning algorithm is developed for handling the problems of pattern classification. The proposed network is consisted of three layers and two stages of learning algorithm and experimented on diabetes dataset [20]. The outcomes are evaluated with batch and online spiking neuron. From the result, it was identified that proposed growing-pruning spiking neural network achieved better accuracy of 71.1%.

Data mining methods like logistic regression and artificial neural networkswith feature selection methods like forward selection and backward elimination are applied on diabetes dataset based on the entropy evaluation method [21]. The experiment was conducted using WEKA.

From the result it was identified that the neural network with backward elimination using percentage split achieved an accuracy of 78.90%. Data mining techniques like logistic regression and artificial neural network are applied on diabetes dataset with feature selection methods like forward selection and backward elimination based on the mean value of the attributes [22]. From the experiment result it was found that an accuracy of 80.46% is achieved by logistic regression when compared to neural network.Using the threshold value of each attribute an experiment was conducted on diabetes dataset and the performances of the data mining methods like logistic regression and artificial neural networks are evaluated with feature selection methods. From the result it was identified that logistic regression using backward elimination achieved an of accuracy of 82.81% when compared to neural network [23]. Using neural network with back propagation and different data mining techniques like J48, naïve bayes and SVM are applied on diabetes dataset to predict the presence or absence of diabetes in a person. A 5-fold cross validation sample is used to improve the performance of the model. Based on the experimental result conducted an accuracy of 83.11% was achieved by back propagation algorithm [24]. In medical field to exploit the patients information, classification systems are widely used on diabetes dataset. The naïve bayes is applied for classification and for attribute selection genetic algorithm is used [25]. From the experimental results an accuracy of 78.69% is achieved. Popular techniques like deep neural networks and SVM are used to identify the presence or absence of diabetes based on the accuracy of cross validation sample on diabetes dataset. An accuracy of 77.86% was achieved from the said method [26]. Six different machine learning algorithms such as J48, mulilayer perceptron, Hoeffding tree, JRip, BayesNet and Random forest are applied on diabetes dataset to evaluate whether the patient is affected or not affected by diabetes. From the result it was found that an accuracy of 77% was achieved from Hoeffding tree [27]. Systematic efforts is done in designing a framework which results in the prediction of diabetes. The author used 3 machine learning classification algorithms on various measures. Experiments are performed on Pima Indians Diabetes dataset. Experimental results determine the adequacy of the designed system with an achieved accuracy of 76.30 % using the Naive Bayes classification algorithm [28]. From the literature survey we can identify that many techniques are applied on the diabetes dataset. Some techniques using full set of attributes and some uses the subsets of the attributes. The classification accuracy achieved is not satisfactory and it can be further improved. In this research work we try to improve the accuracy by using the machine learning techniques like RF, ANN, LR, SVM and NN with k-10fold CV withFSMs methods like forward selection and backward eliminationusingpercentage split as test option.

III. PROPOSED FRAMEWORK

The proposed framework for the research carried out is shown in Figure 1. This consists of the following steps:

1. The PIMA Indian diabetes dataset is selected from the UCI machine learning repository.
2. After selecting the dataset the next step is to do the preprocessing to check for any missing values.
3. For the preprocessed data the entropy value of each attribute is found using equation 1 given below:

$$Info\ (D) = \sum_{i=1}^{m-1} p_i \log_2 (p_i) \quad (1)$$

Where D is the attribute.
i is the attribute index.
p_i is the probability that an attribute in D belongs to
a class.m is the total count of attributes.

4. Based on the entropy value of each attribute apply the FSMs like forward selection and backward elimination. This results in different combinations of attributes.
5. For each combination obtained we predict the CA of the different machine learning techniques using percentage split as test option.
6. The combinationof attributes which gives the better CA is considered as the important attributesto predict the presence of diabetes and the technique with which the better CA is achieved is considered as the best technique.

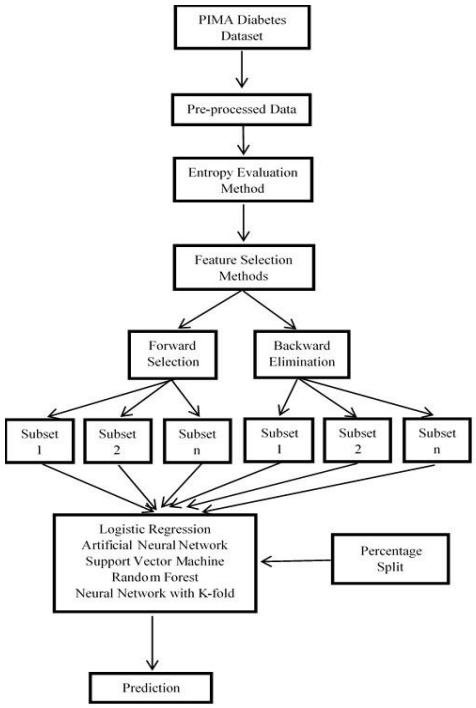

Fig. 1. Proposed Model for Prediction

IV. DATASETS CONSIDERED IN THE PROPOSED WORK

In this proposed work, we have selected the real world PIMA Indian diabetes datasets in Attribute Relation File Format (ARFF). The specification of the datasets is given below in Table I and the attributes with their meaning is given in Table II.

TABLE I. DATASET CONSIDERED FOR THE PROPOSED WORK

Sl. No	MedicalDatasets	No. ofInstances	Total Numberof Attributes	No. ofClasses
1	PIMA Diabetes	768	9	2

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

TABLE II. DATASET ATTRIBUTES AND THEIR MEANING

Attribute	Meaning
preg	Number of times pregnant
plas	Plasma glucose concentration
pres	Diastolic blood pressure (mm Hg)
skin	Triceps skin fold thickness (mm)
insu	2-Hour serum insulin
BMI	Body mass index
pedi	Diabetes pedigree function
Age	Age (years)
outcome	Class variable (0 or 1)

V. RESULTS AND DISCUSSION

Based on the entropy value of each attribute we apply the FSMs like forward selection and backward elimination which results in different combinations of attributes. The different combinations of attributes obtained for forward selection and backward elimination is shown in Table III and Table IV respectively.

TABLE III. ATTRIBUTE COMBINATIONS FROM FORWARD SELECTION

Subset No.	Subset of Attributes	No.of Attributes
1	pres, outcome	2
2	pres, pedi, outcome	3
3	preg, pres, pedi, outcome	4
4	preg, pres, skin, pedi, outcome	5
5	preg, pres, skin, insu, pedi, outcome	6
6	preg, pres, skin, insu, pedi, age, outcome	7
7	preg, pres, skin, insu, mass, pedi, age, outcome	8

TABLE IV. ATTRIBUTE COMBINATIONS FROMBACKWARD ELIMINATION

Subset No.	Subset of Attributes	No. of Attributes
1	preg, plas, skin, insu, mass, pedi, age, outcome	8
2	preg, plas, skin, insu, mass, age, outcome	7
3	plas, skin, insu, mass, age, outcome	6
4	plas, insu, mass, age, outcome	5
5	plas, mass, age, outcome	4
6	plas, mass, outcome	3
7	plas, outcome	2

The CA achieved by different techniques for full set of attributes is shown in Table V. From the Table V we can see that the best CA of 84.52% is achieved by NN with 10-fold CV.

TABLE V: CA ACHIEVED FOR FULL SET OF ATTRIBUTES

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
LR	83.8	83.2	82	81.2	79.9
NN	83.29	83.11	82.88	82.8	83.4
NN with 10 Fold	83.70	**84.52**	83.35	83.55	83.8
SVM	72.9	69.9	64.7	66.4	65.3
RF	83.8	83	83.7	84.2	84.5

The CA achieved for different subsets of Table III is shown from Table VI through Table XII.

TABLE VI: CA ACHIEVED FOR SUBSET NO. 1 OF TABLE III (PRES ANDOUTCOME)

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
LR	59.2	59.5	60.6	61.2	58.2
NN	77.6	77.61	77.7	78.13	78.17
NN with 10 Fold	77.13	77.62	77.23	77.7	77.56
SVM	50.9	51	51.2	51.4	50.3
RF	53.7	57.4	57.2	54.1	54

TABLE VII: CA ACHIEVED FOR SUBSET NO. 2 OF TABLE III (PRES, PEDI and OUTCOME)

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
LR	62.6	62.7	64.3	62.6	63.1
NN	78.51	77.89	77.99	79.0	79.24
NN with 10 Fold	77.49	78.10	77.97	77.60	78.73
SVM	55.6	57.4	58.1	56.7	56.6
RF	53.8	58	58.5	59.9	58.2

TABLE VIII: CA ACHIEVED FOR SUBSET NO. 3 OF TABLE III (PREG, PRES, PEDI and OUTCOME)

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
LR	66.3	67	68.3	67.6	67.6
NN	78.76	78.14	77.55	78.3	78.44
NN with 10 Fold	78.44	79.47	78.58	79.05	78.60
SVM	56.6	56.4	55.1	56.9	55.1
RF	59.1	59	62	61.2	62.9

TABLE IX: CA ACHIEVED FOR SUBSET NO. 4 OF TABLE III (PREG, PRES, SKIN, PEDI and OUTCOME)

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
LR	68.1	68.5	68.3	69.3	67
NN	79.05	76.98	77.57	78.68	78.84
NN with 10 Fold	78.20	78.82	78.9	78.64	78.94
SVM	57.9	57	59	55	56.5
RF	63.6	59.8	62.9	62.8	63.7

TABLE X: CA ACHIEVED FOR SUBSET NO. 5 OF TABLE III (PREG, PRES, SKIN, INSU, PEDI and OUTCOME)

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
LR	69.2	71.7	71.7	71.4	73.1
NN	78.8	78.36	76.68	79.24	79.25
NN with 10 Fold	78.3	78.67	79.19	78.22	78.19
SVM	59	58.4	56.8	59	61.7
RF	67.5	64.7	68	62.4	67.3

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

TABLE XI: CA ACHIEVED FOR SUBSET NO. 6 OF TABLE III (PREG, PRES, SKIN, INSU, PEDI, AGE and OUTCOME)

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
LR	69.2	71.9	70.3	72.9	70.4
NN	78.95	79.06	79.63	79.38	79.11
NN with 10 Fold	79.50	79.15	79.76	79.75	79.93
SVM	61.1	58.5	61.8	55.8	59.1
RF	70.6	71.9	73.2	73.1	68.5

TABLE XII: CA ACHIEVED FOR SUBSET NO. 7 OF TABLE III (PREG, PRES, SKIN, INSU, BMI, PEDI, AGE and OUTCOME)

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
LR	73.2	70.5	69.4	67.4	66.2
NN	80.62	80.29	80.87	80.92	81.88
NN with 10 Fold	80.90	80.82	80.93	80.67	80.80
SVM	60	62.5	56.6	58.3	59.4
RF	73.6	75.5	72.7	71.9	70.3

The CA achieved for different subsets of Table IV is shown from Table XIII through Table IXX.

TABLE XIII: CA ACHIEVED FOR SUBSET NO. 1 OF TABLE IV (PREG, PLAS, SKIN, INSU, BMI, PEDI, AGE and OUTCOME)

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
LR	84.3	83.1	83.5	83	82.2
NN	83.11	83.08	83.02	83.12	83.96
NN with 10 Fold	83.66	84.09	83.21	84.14	84.41
SVM	70.9	70	70.9	72.8	66.7
RF	82.8	83.4	84.1	83.3	83.6

TABLE XIV: CA ACHIEVED FOR SUBSET NO. 2 OF TABLE IV (PREG, PLAS, SKIN, INSU, BMI, AGE and OUTCOME)

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
LR	83.7	83.7	82.7	82.3	82.5
NN	82.70	82.48	82.30	82.29	82.70
NN with 10 Fold	83.73	83.78	84.06	83.99	**85.24**
SVM	69.7	72.2	69.6	70.9	70.5
RF	81.4	83.1	82.8	81.5	83.2

TABLE XV: CA ACHIEVED FOR SUBSET NO. 3 OF TABLE IV (PLAS, SKIN, INSU, BMI, AGE and OUTCOME)

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
LR	82.6	82.1	80.1	80.4	79.31
NN	82.69	83.10	82.87	83.26	83.58
NN with 10 Fold	83.01	82.92	84.13	83.55	83.02
SVM	68.9	69.4	71	72	69.5
RF	81.1	81.4	81	83.4	81.7

TABLE XVI: CA ACHIEVED FOR SUBSET NO. 4 OF TABLE IV (PLAS, INSU, BMI, AGE and OUTCOME)

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
LR	82.4	81.3	80.6	81.1	80.1
NN	82.80	83.11	82.87	83.25	83.63
NN with 10 Fold	82.92	83.48	83.77	82.57	83.69
SVM	72.6	70.5	70	72.4	70.4
RF	81.7	81.6	81.1	82.9	80.3

TABLE XVII: CA ACHIEVED FOR SUBSET NO. 5 OF TABLE IV (PLAS, BMI, AGE and OUTCOME)

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
LR	83.7	82.2	81.1	81.1	80.1
NN	82.76	83.23	82.96	83.43	83.95
NN with 10 Fold	83.50	83.63	83.52	84.15	82.87
SVM	72	69.8	69.7	73.1	72.7
LR	83.7	82.2	81.1	81.1	80.1

TABLE XVIII: CA ACHIEVED FOR SUBSET NO. 6 OF TABLE IV (PLAS, BMI and OUTCOME)

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
LR	83.5	81.7	80.2	80.7	80.4
NN	82.75	82.96	82.84	83.22	83.66
NN with 10 Fold	83.30	83.14	82.70	83.56	83.69
SVM	68.1	68.7	68.5	67.6	66.9
RF	74.3	74.6	74.2	74.2	75.2

TABLE IXX: CA ACHIEVED FOR SUBSET NO. 7 OF TABLE IV (PLAS, BMI and CLASS)

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
LR	82.1	81.6	82	81.3	80.5
NN	81.94	82.18	81.99	82.67	82.88
NN with 10 Fold	82.65	82.61	82.06	82.41	82.88
SVM	65.3	65.7	65.8	66.6	68.8
RF	71.3	68.6	67.7	70.3	67.7

The Comparison of the CA achieved by different methods and the proposed machine learning technique is shown in Figure 2.

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

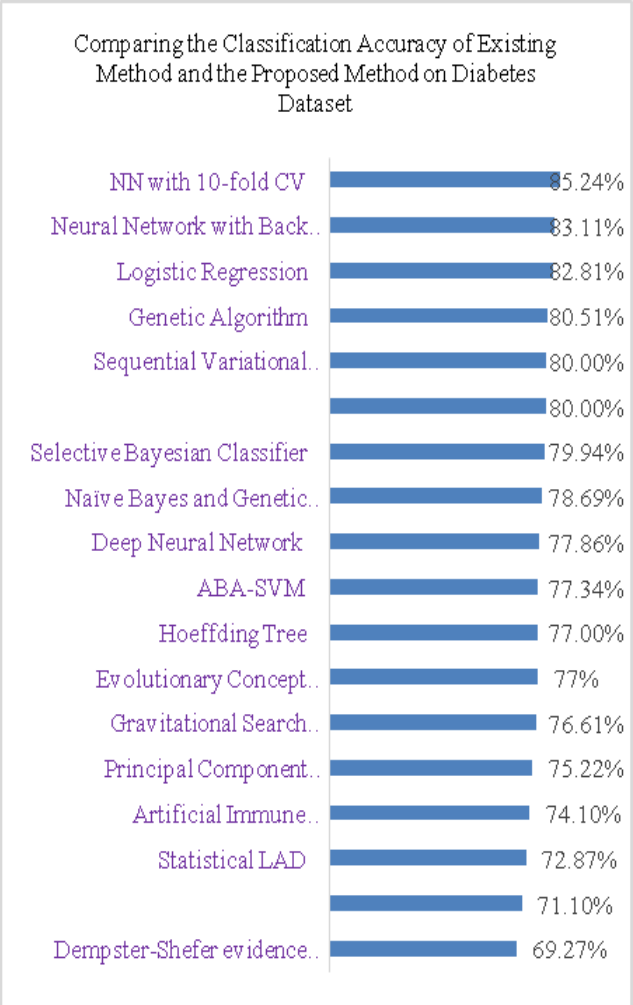

Fig. 2. Comparison of CA of various methods with the proposed machine learning technique.

From the above figure CA achieved for different subsets of attributes obtained by forward selection and backward elimination we can see that the best CA of 85.24% is achieved by NN with 10-fold and is shown in Table XIV with only 6 attributes.

VI. CONCLUSION

In the proposed research work we apply machine learning techniques like LR, SVM, RF, ANN and NN with 10-fold CV with FSMs like forward selection and backward elimination on diabetes dataset with percentage split as test options. From the experimental results it is identified that NN with 10-fold CV achieves CA of **85.24%** with only 6 attributes. The result achieved is better than the other machine learning techniques considered for the experimentation and also better than the other methods discussed in literature review and full set of attributes which is **84.52%** and also we state that diabetes patients need not go for all tests to predict diabetes instead important features or tests what we have identified through research.

REFERENCES

1. S. K. Wasan and V. Bhatnagar and H. Kaur, "The Impact of Data Mining Techniques on Medical Diagnostics," Data Science Journal, vol. 5, pp.119-126, 2006.
2. M. K. Bodla and S. M. Malik and M. T. Rasheed and M. Numan and M. Z. Ali and J. B. Brima,"Logistic Regression and Feature Extraction Based Fault Diagnosis of Main Bearing of Wind Turbines," IEEE 11th International Conference on Industrial Electronics and Applications (ICIEA), pp. 1628-1633, 2016.
3. C. P. Prathibhamol and K. V. Jyothy and B. Noora,"Multi Label Classification Based on Logistic Regression (MLC-LR)," International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 2708-2712, 2016.
4. B. K. Raghavendraand J. B. Simha,"Performance Evaluation of Logistic Regression and Neural Network Model with Feature Selection Methods and Sensitivity Analysis on Medical Data Mining," International Journal of Advanced Engineering Technology, vol. 2, no. 1, pp. 289-298, 2011.
5. C. Cortes and V. Vapnik, "Support Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.
6. T. K. Ho, "The Random Subspace Method for Constructing Decision Forests," IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 20, no. 8, pp. 832-844, 1998.
7. W. G. Touw and J. R. Bayjanov and L. Overmars and L. Backus and J. Boekhorst and M. Wels and S. A. F. T. V. Hijum, "Data Mining in the Life Sciences with Random Forest: a Walk in the park or lost in Jungle?," Briefings in Bioinformatics, vol. 14, no. 3, pp. 315-326, 2012.
8. M. L. Raymer and T. E. Doom and L. A. Kuhn and W. F. Punch, "Knowledge Discovery in Medical and Biological Datasets Using a Hybrid Bayes Classifier/Evolutionary Algorithm,"IEEE Transactions on Systems, Man and Cybernatics, vol. 33, no. 5, pp. 802-813, 2003.
9. M. B. Dowlatshahi and M. Rezaeian, "Training Spiking Neurons with Gravitational Search Algorithm for Data Classification," IEEE 1st International Conference on Swarm Intelligence and Evolutionary Computation, pp. 53-58, 2016.
10. E. Tuba and M. Tuba and D. Simian,"Adjusted Bat Algorithm for Tuning of Support Vector Machine Parameters," IEEE Congress on Evolutionary Computation, pp. 2225-2232, 2016.
11. R. Bruni and G. Bianchi, "Effective Classification Using a Small Training Set Based on Discretization and Statistical Analysis," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 9, pp. 2349-2361, 2015.
12. P. Sykacek and S. Roberts, "Adaptive Classification by Variational Kalman Filtering," Advances in Neural Information Processing Systems (NIPS), pp. 737-744, 2002.
13. F. J. Li and Y. H. Qian and J. T. Wang and J. Y. Liang, "Multigranulation Information Fusion: A Dempster-Shafer Evidence Theory Based Clustering Ensemble Method," Proceedings of IEEE International Conference on Machine Learning and Cybernetics (ICMLC), vol. 1, pp. 58-63, 2015.
14. S. J. Perantonis and V. Virvilis, "Input Feature Extraction for Multilayered Perceptrons Using Supervised Principal Component Analysis," Neural Processing Letters, vol. 10, no. 3, pp. 243-252, 1999.
15. C. A. Ratanamahatana and D. Gunopulos, "Feature Selection for the Naïve Bayesian Classifier Using Decision Trees," Applied Artificial Intelligence (AAI), vol. 17, no. 5-6, pp. 475-487, 2003.
16. F. Divina and E. Marchiori, "Knowledge-Based Evolutionary Search for Inductive Concept Learning," Knowledge Incorporation in Evolutionary Computation, Springer, vol. 167, pp. 237-253, 2005.
17. G. T. Hilda and R. R. Rajalaxmi, "Effective Feature Selection for Supervised Learning Using Genetic Algorithm," IEEE 2nd International Conference on Electronics and Communication Systems (ICECS 2015), pp. 909-914, 2015.
18. Blayvas and R. Kimmel, "Machine Learning via Multiresolution Approximations," IEICE Transaction on Information System, vol. E86-D, no. 7, pp. 1172-1180, 2003.
19. Watkins and J. Timmis and L. Boggess, "Artificial Immune Recognition System (AIRS): An Immune Inspired Supervised Learning Algorithm," Genetic Programming and Evolvable Machines, vol. 5, no. 3, pp. 291-317, 2004.
20. S. Dora and S. Sundaram and N. Sundararajan,"A Two Stage Learning Algorithm for a Growing-Pruning Spiking Neural Network for Pattern Classification Problems,"International Joint Conference on Neural Networks (IJCNN), pp. 1-7, 2015.
21. S. Raghavendra and M. Indiramma, "Performance Evaluation of Logistic Regression and Artificial Neural Network Model with Feature Selection Methods Using Cross Validation Sample and Percentage Split on Medical Datasets," International Conference on Emerging Research in Computing, Information, Communication and Applications, vol. 2, 2014.

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

22. S. Raghavendra and M. Indiramma, "Classification and Prediction Model Using Hybrid Technique for Medical Datasets," International Journal of Computer Applications, vol. 127, no. 5, pp. 20-15, 2015.
23. S. Raghavendra and M. Indiramma, "Hybrid Data Mining Model for the Classification and Prediction of Medical Datasets," International Journal of Knowledge Engineering and Soft Data Paradigms," vol. 5, no. 3/4, pp. 262- 284, 2017.
24. F. G. Woldemichael and S. Menaria, "Prediction of Diabetes Using Data Mining Techniques," 2nd International Conference on Trends In Electronics and Informatics, pp. 414-418 , 2018.
25. D. K. Choubey and S. Paul and S. Kumar and S. Kumar,"Classification of Pima Indian Diabetes Dataset Using Naïve Bayes with Genetic Algorithm as an Attribute Selection," Communication and Computing Systems, Taylor & Francis Group, pp. 451-455, 2017.
26. S. Wei and X. Zhao and C. Miao, "A Comprehensive Exploration to the Machine Learning Technique for Diabetes Dataset," IEEE 4th World Forum on Internet of Things, pp. 291-295, 2018.
27. F. Mercaldo and V. Nardone and A. Santone,"Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques,"Procedia Computer Science,vol. 112, pp. 2519–2528, 2017.
28. D. Sisodia and D. S. Sisodia."Prediction of Diabetes using Classification Algorithms,"Procedia Computer Science, vol. 132, pp. 1578–1585, 2018.

AUTHORS PROFILE

Dr. Raghavendra S. is currently working as Associate Professor in the Department of Computer Science and Engineering at CHRIST DEEMED TO BE UNIVERSITY, Bangalore. He completed his Ph.D. degree in Computer Science and Engineering from VTU, Belgaum, India in 2017 and has 14 years of teaching experience. His interests include Data Mining and Big data.

Santosh Kumar J.is currently working as Associate Professor in the Department of Computer Science and Engineering at K.S.School of Engineering and Management, Bangalore. He is pursuing Ph.D. in VTU, Belgaum, India. He has 10 years of teaching and 3 years of industry experience. He is specialized in Big data streaming analysis. His reserach topics includesBig data with machine learning.

Dr. Raghavendra B.K. Pursued P.hd From VTU Belagavi Karnataka and Masters from VTU Belagavi and Bachelors from Bengaluru University Karnataka He published nearly 15 reputed journals and His Area of interest is Data mining and Big data He currently working in KSSEM Bengaluru.

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