International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

Type II Diabetes Prediction Using Combo of SVM ANN and Random Tree

Minyechil Alehegn, Rahul Raghvendra Joshi

Abstract: ***In 21th century, IT plays a very important and helpful role in health care industries acting as a savior to human life. Data mining and machine learning are two sides of healthcare-IT. Proposed system considers one of the most common chronic diseases called diabetes. India and almost all other countries are worried about diabetic patients, so diabetes can termed as a global chronic disease. In this paper, well-known predictive machine learning techniques viz. SVM, Random Tree and ANN are applied on PIMA dataset. Results of SVM, ANN, and RT are 90.1%, 88.02%, and 83.59% respectively.***

Index Terms: PIMA, SVM, ANN, Random Tree etc.

I. INTRODUCTION

Our world faces different challenges and problem in different seasons. These problems occurred in civilized as well as uncivilized country. They are not limited only for war, natural disaster, it also includes different dangerous health problem. These diseases based on their level and some criteria kill human being at early stage. The one which are grouped under dangerous diseases is also called as diabetes. Diabetes is a dangerous now days. This risky disease cut human life at early stage, as a human being we have to help other human beings from this disease. There are different health care services which help patients suffering with diabetes. Information Technology (IT) has its own role to solve different human being problems including diseases related problems by applying different technologies starting from old to latest one. In health related problems machine learning can play a great role as it is one of the branches of artificial intelligence. Different ML techniques help in the prediction purpose to predict the current, previous as well the future state of a specific disease. These dangerous diseases can cause harm to various or different human body parts, such as eye, nerve system, heart and kidney. In diabetes mellitus (DM) patient will not get time to recover if it reaches high sugar level as it was not normal or through formal medicine recovery will not be possible. According to the IDF (International Diabetes federation) report on 14 November 2017, now-a-day's or currently there are more than 199 million women are suffering from DM [17]. They predicted that by 2040 this number will increase to 313 million. The report indicated that diabetes is the ninth level cause for death. Every year 2.1 million people die due to DM. 1 out 7 births was reported with diabetes.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Minyechil Alehegn*, School of Computing and Informatics, Department of IT, MizanTepi University, Tepi, Ethiopia.
Rahul Raghvenrda Joshi, CS/IT, Symbiosis International (Deemed) University/ Symbiosis Institute of Technology, Pune, Maharashtra India.

The report indicated that most of the patient are live with Type 2 diabetes also and it is caused due to modern life style [17]. So, diabetes is a considerable chronic disease now-a-days.

Considering these problem, some machine learning techniques need to be proposed to address such problems. The proposed method or way follows various Data mining (DM) or machine learning mechanisms (MLMs) in order to predict diabetes at an early stage to save human life. The algorithm proposed here includes SVM, Artificial Neural Network (ANN), and Random Tree to prediction diabetes.

II. RELATED WORK

Ioannis et al. [1] used data mining and machine learning algorithms as key tactics against huge volume of diabetes related data for retrieving novel knowledge about it. Feature selection applied by Yawen et al. [2] is an ensemble approach using five different classification algorithms such as KNN, SVM.DT, GBDT and RF. These algorithms were applied on three different datasets. Ensemble approach showed better accuracy than other classification techniques. Tao et al. [3] studied type II diabetes using most commonly used data mining techniques with feature selection mechanism by creating a new frame. Sajida et al. [4] proposed machine learning classification techniques were not analyzed using the evaluation techniques like cross validation. The algorithms were applied on four different global dataset which are available online viz. diabetes; nutrition, ecoli protein, and mushroom were used. For diabetes dataset Naïve Bayes (NB) provided good accuracy of 77.01 % compared to other data mining algorithms. On contrary ANN and KNN proved good for nutrition, mushroom, and ecoli protein datasets. Anjli and varun [5] used binary classification method (BCM) which was applied on two different diabetes datasets which are available online. Evaluation method for 10K cross validation method applied in multiple iterations. Ranker method (RM) used for feature selection. Herbert et al. [6] concluded that HbA1c (glycated haemoglobin) is one of diabetes causing factor. Farhi et al [7] used data mining to retrieve novel knowledge which was not known before. Seokho et al [8] used svm machine learning algorithm for disease analysis. If correct treatment provided to diabetic patients then their early recovery could be possible. J. Pradeep and S. Balamurali [9] did comparison of various algorithms by removing noisy data and after data cleansing four algorithm were applied. Random Forest, Decision Tree (J48), KNN, and SVM applied. Before removal of noisy data accuracy was 73.8 %, J48 shows better accuracy in comparison with others. Nongyao and Rungruttikarn [10] developed a web application for disease analysis.

Four well known techniques RF, ANN, Logistic Regression and Naïve Bayes applied. After the examination of these four algorithms boosting and bagging technique were applied. Random Forest gives better than other classification method. Ayush and Divya [11] used classification and regression tree (CART) applied with 10 cross validation evaluation method. Day to day life style is has relationship with diabetes. CART showed accuracy of 75%. Blood Pressure is the significant factor caused due to eating habits. C. Kalaiselvi and Nasira [12] an abnormal blood glucose level has the probability of leading to heart and cancer. A NFIS (Nuero Fuzzy Inference System) was applied for prediction and provided 80% disease prediction accuracy. Abdullah et al [13] analyzed predictive analysis of diabetic treatment using SVM. The data set was world health organization report (WHOR). Oracle Data miner (ODM) tool was used for analysis. Xue-Hui et al [14] used logistic regression, J48, and artificial neural network (ANN), applied on real world dataset through distributing questioner. J48 showed accuracy of 78.27 % than other data mining classification techniques (DMCTs). Asma [15] applied decision tree was using WEKA. 78.1768% accuracy she got using j48 prediction technique. Emirhan et al [16] concluded that by providing suitable dose of medicine to diabetic patients, it is possible to save life in early stage. Rough Set and ANFIS applied using WEKA. ANFIS outperformed than Rough Set. If two patterns are close to each other in terms of disease parameters, then third patient's pattern will be taken for analysis.

III. PROPOSED PREDICTION AND CLASSIFICATION METHOD

Pima Indian Diabetes Dataset (PIDD) has following attributes.

Table 1 PIDD attributes

	Attribute Name	Description
1	Pregnancies	The frequency of become pregnant
2	Glucose	Plasma glucose concentration - a 2 Hrs. in oral glucose tolerance test
3	BP	The blood pressure of the patient
4	Skin Thickness	The thickness of the patient skin
5	Insulin	It is 2 Hr. serum insulin
6	BMI	Body mass index which can be obtained from weight and length of the patient
7	Diabetes Pedigree Function	Family history
8	Age	Patient's age
	Class	Presence or absence

3.1 SVM

This algorithm is well known for prediction. It is also called binary algorithm. It is used for binary like presence or absence or diabetic or non-diabetic etc.
SVM follows following steps:

1. Identification of right hyper plane
2. Maximizing the spaces between neighbor data point
3. Adding a feature

Z=X^2+Y^2. SVM solves such problem.

4. Apply a binary classifier.

3.2 ARTIFICIAL NEURAL NETWORK (ANN)

ANN is composed of different nodes. It is also defined as emulation for biological neural network system. It contains input, hidden and output layer. The algorithm has following steps:

1. Initialize the weights randomly
2. Use learning algorithm
3. Set of different training examples
4. Examples encoded as inputs
5. Conversion of output to meaning full outcome

3.3 RANDOM TREE

RT (Random tree) is a group of different decision trees that indicates that random tree works as follow the decision tree operator except, for each divided, only a random subset of features. A RT is a tree exhausted at a chance from the group of successful or achievable trees.
Algorithm:

1. RFT (D, DPT, L)
2. FOR $L = 1 \ldots L$ DO
3. $T(L) \leftarrow CBT($COMPLETE BINARY TREE(CBT)) OF
4. DEPTH DPT WITH RFS(RANDOM FEATURE SPLITS)
5. $F(L) \leftarrow$ THE FUNCTION COMPUTED BY $T(L)$, WITH LEAVES FILLED IN BY D
6. END FOR
7. RETURN $F(X\hat{}) = \text{SGN} \sum K F(\text{K})(X\hat{})$

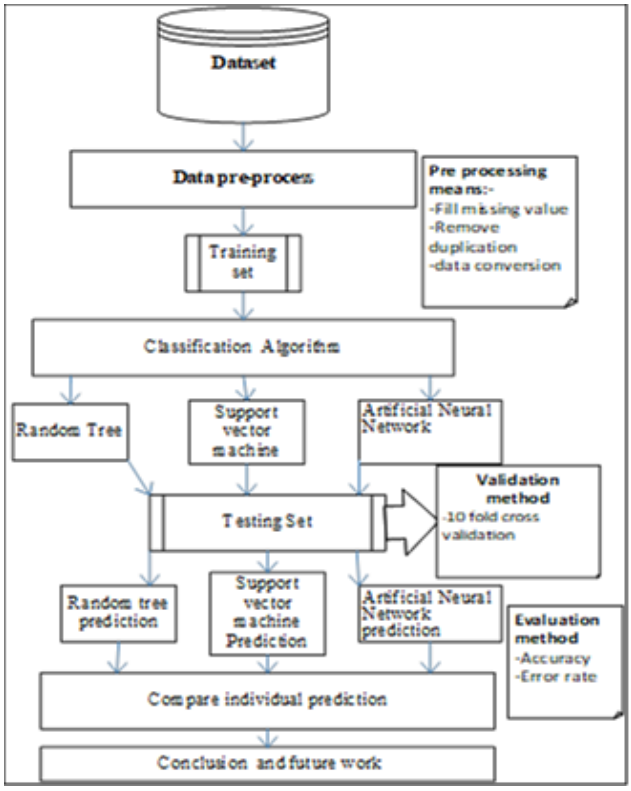


Fig 1Proposed Work Flow

IV. RESULT

$$\text{Acuracy} = 100 * \left(\frac{\text{correctly claccified}}{\text{correctly classified} + \text{incorrectly classified}} \right)$$

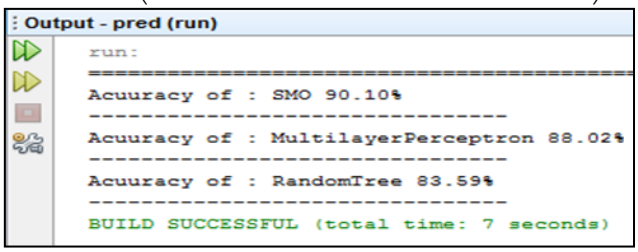


Table 2: The predictive accuracy considered algorithms

Algorithm	Accuracy for correctly classified instances	Accuracy of incorrectly classified instances
SVM	90.1%	9.9%
Random tree(RT)	83.59	16.41
ANN	88.02	11.98

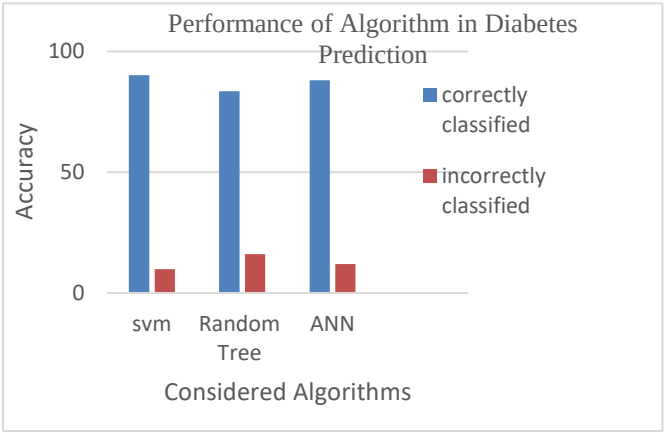
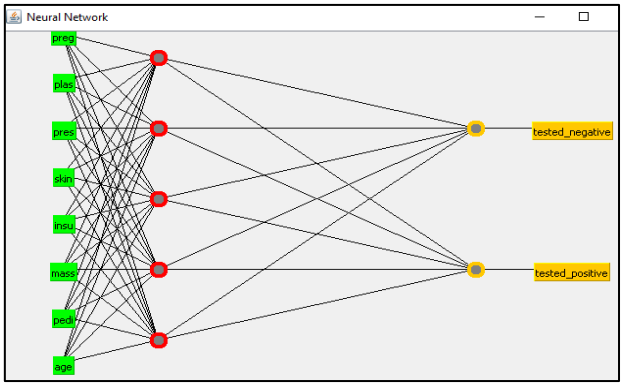


Fig 2: Accuracy of algorithms



Input layer Hidden layer Output layer

Fig 3: ANN for PIDD

Input layer helps to feed data to the neural network. The attributes of diabetes are preg, plas, pres skin, insulin, mass (body mass index), pedi and age are basically input to ANN. Each data can be feed to given neurons, the middle layer used as the intermediary between the input layer and output one. Output neurons output the generated forecasts. Output prediction provides two results namely tested negative and tested positive or diabetic or non diabetic.

V. CONCLUSION AND FUTURE WORK

In past various Machine learning and data mining techniques and their applications were reviewed in different perspectives. Data mining and machine learning applied on different medical datasets including diabetes dataset too. Data mining and machine learning techniques have varying processing powers or prediction values for different data sets. In proposed system, PIDD dataset available on UCI repository containing 768 records with 8 attributes is analyzed. 10K cross validation for both in single and multiple iterations with 90 % for training and 10 % for testing considered. Proposed system used known and commonly used machine learning algorithm including deep learning algorithm also. Out of considered three techniques SVM showed better performance with accuracy of 90.1%. The proposed system need to be tested with large datasets in future and or by hybridizing these algorithms in order to increase their prediction.

REFERENCES

1. Kavakiotis, Ioannis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, and Ioanna Chouvarda. "Machine learning and data mining methods in diabetes research." Computational and structural biotechnology journal (2017).
2. Xiao, Yawen, Jun Wu, Zongli Lin, and Xiaodong Zhao. "A deep learning-based multi-model ensemble method for cancer prediction." Computer methods and programs in biomedicine153 (2018): 1-9.
3. Zheng, Tao, Wei Xie, Liling Xu, Xiaoying He, Ya Zhang, Mingrong You, Gong Yang, and You Chen. "A machine learning-based framework to identify type 2 diabetes through electronic health records." International journal of medical informatics 97 (2017): 120-127.
4. Perveen, Sajida, Muhammad Shahbaz, Aziz Guergachi, and Karim Keshavjee. "Performance analysis of data mining classification techniques to predict diabetes." Procedia Computer Science 82 (2016): 115-121.
5. Negi, Anjli, and Varun Jaiswal. "A first attempt to develop a diabetes prediction method based on different global datasets." In Parallel, Distributed and Grid Computing (PDGC), 2016 Fourth International Conference on, pp. 237-241. IEEE, 2016.
6. Jelinek, Herbert F., Andrew Stranieri, Andrew Yatsko, and Sitalakshmi Venkatraman. "Data analytics identify glycated haemoglobin co-markers for type 2 diabetes mellitus diagnosis." Computers in biology and medicine 75 (2016): 90-97.
7. Marir, Farhi, Huwida Said, and Feras Al-Obeidat. "Mining the Web and Literature to Discover New Knowledge about Diabetes." Procedia Computer Science 83 (2016): 1256-1261.
8. Kang, Seokho, Pilsung Kang, Taehoon Ko, Sungzoon Cho, Su-jin Rhee, and Kyung-Sang Yu. "An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction." Expert Systems with Applications 42, no. 9 (2015): 4265-4273.
9. Kandhasamy, J. Pradeep, and S. Balamurali. "Performance analysis of classifier models to predict diabetes mellitus." Procedia Computer Science 47 (2015): 45-51.
10. Nai-arun, Nongyao, and RungruttikarnMoungmai. "Comparison of classifiers for the risk of diabetes prediction." Procedia Computer Science 69 (2015): 132-142.
11. Anand, Ayush, and Divya Shakti. "Prediction of diabetes based on personal lifestyle indicators." In Next Generation Computing Technologies (NGCT), 2015 1st International Conference on, pp. 673-676. IEEE, 2015.
12. Kalaiselvi, C., and G. M. Nasira. "A new approach for diagnosis of diabetes and prediction of cancer using ANFIS." In Computing and Communication Technologies (WCCCT), 2014 World Congress on, pp. 188-190. IEEE, 2014.
13. Aljumah, Abdullah A., Mohammed Gulam Ahamad, and Mohammad Khubeb Siddiqui. "Application of data mining: Diabetes health care in young and old patients." Journal of King Saud University-Computer and Information Sciences 25, no. 2 (2013): 127-136.

14. Meng, Xue-Hui, Yi-Xiang Huang, Dong-Ping Rao, Qiu Zhang, and Qing Liu. "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors." The Kaohsiung journal of medical sciences 29, no. 2 (2013): 93-99.
15. Al Jarullah, Asma A. "Decision tree discovery for the diagnosis of type II diabetes." In Innovations in Information Technology (IIT), 2011 International Conference on, pp. 303-307. IEEE, 2011.
16. Yıldırım, EmirhanGülçin, AdemKarahoca, and Tamer Uçar. "Dosage planning for diabetes patients using data mining methods." Procedia Computer Science 3 (2011): 1374-1380.
17. https://www.idf.org/our-activities/world-diabetes-day/wdd-2017.html
18. https://www.kaggle.com/uciml/pima-indians-diabetes-database.

AUTHORS PROFILE

Mr. Minyechil Alehegn is lecture in department of Information Technology at Mizan Tepi University, Ethiopia. He received M.Tech in Computer Science and Engineering from Symbiosis International University, Pune, Maharashtra, India. He received BSc degree in IT from wollega university, Ethiopia in 2014. His research interests include IoT, AI, Machine Learning, Networking and Big Data.

Mr. Rahul Rghavendra Joshi is currently working as an assistant professor in CS/IT department at Symbiosis Institute of Technology, Pune, Maharashtra, India. He has authored more than 40 papers in reputed indexed journals. His research interests include Machine Learning, IoMT and Distributed Incremental Clustering. He is pursuing PhD under the guidance of Dr. Preeti Mulay at Symbiosis International (Deemed) University, Pune, Maharashtra, India.