A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques

Amani Yahyaoui
Department of Sotware Engineering
Istanbul Sabahattin Zaim University,
Istanbul, Turkey
amani.yahyaoui@izu.edu.tr

Akhtar Jamil
Department of Computer Engineering
Istanbul Sabahattin Zaim University,
Istanbul, Turkey
https://orcid.org/0000-0002-2592-1039

Jawad Rasheed
Department of Computer Engineering
Istanbul Sabahattin Zaim University,
Istanbul, Turkey
jawad.rasheed@izu.edu.tr

Mirsat Yesiltepe
Dept. of Mathematical Engineering,
Yildiz Technical University,
Istanbul, Turkey
mirsaty@yildiz.edu.tr

***Abstract*— With the continuing increase in the number of the deadly diseases that threaten both human health and life, medical Decision Support Systems (DSS) continue to prove their effectiveness in providing physicians and other healthcare professionals with support in clinical decision making. Among these dangerous diseases, diabetes continues to be one of the leading one that has caused several deaths in the world. It is characterized by an increase in blood sugar levels which can have severe effects on other human organs. According to the International Diabetes Federation (IDA), 382 million people are living with diabetes and by 2035, these statistics will double to reach 592 million. In this paper, we propose a DSS for diabetes prediction based on Machine Learning (ML) techniques. We compared conventional machine learning with deep learning approaches. For conventional machine learning method, we considered the most commonly used classifiers: Support Vector Machine (SVM) and the Random Forest(RF). On the other hand, for Deep Learning (DL) we employed a fully Convolutional Neural Network (CNN) to predict and detect the diabetes patients. The proposed system is evaluated on publicly available Pima Indians Diabetes database which consisted of total 768 samples each with 8 features. 500 samples were labeled as non-diabetic while 268 were diabetic patients. The overall accuracy obtained using DL, SVM and RF was 76.81%, 65.38% and 83.67% respectively. The experimental results show that RF was more effective for diabetes prediction compared to deep learning and SVM methods.**

Keywords*— *Decision Support Systems, diabetes, machine learning, deep learning, Support Vector Machine, Random Forest, Convolutional Neural Network.

I. INTRODUCTION

Diabetes mellitus or diabetes is one of the incurable chronic diseases caused by lack or absence of a hormone called insulin [1]. It is an essential hormone produced by the pancreas that allows the cells to absorb glucose (blood sugar) from food supplies in order to provide them the necessary energy [2]. The presence of high blood sugar levels in the blood is known as Hyperglycemia in medical terms. This situation can occur for two main reasons: (1) when the body cannot make insulin required by the blood cells (2) the body cannot respond to insulin properly. The body needs insulin so glucose in the blood can enter the cells of the body where it can be used for energy. However, if the body fails to utilize glucose to produce energy, it builds up in the blood resulting in hyperglycemia. This can cause serious health problems such as diabetic ketoacidosis, nonketotic hyperosmolar, cardiovascular disease, stroke etc.

According to the World Health Organization, diabetes is one of the leading causes of death worldwide and about 422 million people worldwide have diabetes. Indeed, it caused the deaths of 1.6 million people in 2016 [3].

There are two main types of diabetes, type1 and type 2. The diabetes type 1 span 5 to 10% of all diabetes cases. This type of diabetes appears most often during childhood or adolescence and characterized by the partial functioning of pancreas. At the beginning, type 1 diabetes does not develop any symptoms, as the pancreas remains partially functional. The disease only becomes apparent when 80-90% of pancreatic insulin-producing cells are already destroyed [4].

The diabetes type 2 presents 90% of all diabetes cases. This type of diabetes is characterized by chronic hyperglycemia and the body's inability to regulate blood sugar levels, which causes a too high glucose (sugar) level in the blood. This disease usually occurs in older adults and affects more obese or overweight people [5].

In medicine, doctors and current research confirm that if the disease is discovered at an early stage, the chances of recovery will be greater. With the continuous advancement of technology, machine learning and deep learning techniques have become very useful in early prediction and disease analysis. Among these techniques, Support Vector Machine (SVM), the Random Forest (RF) and the Convolutional Neural Network (CNN) are used in this research to predict the diabetes.

Recently, several researches have focused on predicting diabetes using machine learning and deep learning techniques. For instance, in [6], authors have proposed a deep learning-based method for diabetes data classification by using the Deep Neural Network (DNN) method. The proposed system was experimented on Pima Indians Diabetes data set. The proposed system has shown good classification accuracy (86.26%) which shows the effectiveness of the DNN in helping doctors to predict the disease.

In [7], authors have presented a theoretical research based on three classification method from machine learning techniques which are the SVM, the Logistic Regression (LR) and the Artificial Neural Network (ANN).

978-1-7281-3992-0/19/$31.00 ©2019 IEEE

In [8], authors have proposed an hybrid system for diabetes prediction by using the Boltzman method from deep learning techniques to predict whether the patient is diabetic or not and by using the decision tree method from the machine learning techniques to classify either the diabetic patient is having type 1 or type 2 diabetes. The obtained results prove the good performance of the deep learning method by reaching an accuracy of 80% and also for the machine learning method by giving 94%. From this research, we can conclude that deep learning and machine learning have mixed their skills and advantages to give birth to a powerful hybrid system that can absolutely help doctor not only in prediction diabetes but also in classifying the diabetes type.

Another recent research presented in [3] have proposed a hybrid diabetes detection system based on mixing two common techniques from deep learning which are the Long Short Term Memory (LSTM) method and the Convolutional Neural Network (CNN) method. The obtained results confirm the effectiveness of the deep learning by reaching 95.7%.

In addition, in [9], authors also have presented a deep learning approach to identify diabetes by using the Recurrent Deep Neural Network (RDNN) method. This approach was evaluated by using the public well know data set Pima Indians diabetes. The performance of the proposed approach was very good and has reached an accuracy of 81%.

Moreover, in [10], authors have proposed and hybrid system composed by three common algorithms of machine learning which are the Decision Tree (DT), the Neural Network (NN) and the Random Forest (RF) methods to predict the diabetes. The data used was taken from Luzhou hospital, china and composed by 68994 healthy and diabetic patients. In this research, the Principal Component Analysis (PCA) was also used to reduce the data set dimensionality. The obtained accuracy has reached 80% which can be considered as good results.

In this paper, three different learning-based methods are compared for prediction of diabetes disease on the above data set. Our main objective is to analyze the efficiency of conventional machine learning and deep learning approaches for diabetic prediction. We used SVM and RF as part of conventional machine learning approaches and CNN as part of deep learning method. Literature review shows that both SVM and RF have proven to be effective for many classification algorithms. Similarly, CNNs have recently been widely used for many classification and recognition tasks as well. Therefore, we selected these algorithms to evaluate their performance for diabetes detection to develop a decision support system. Extensive experiments were performed on our data set. For each method, same number of training and testing samples were selected. All the analysis and visualization are carried out in python 3.6 within the Anaconda 5 environment.

II. Materials

The proposed method was evaluated on the online available PIMA Indians diabetes dataset which is available online and can be downloaded from UCI machine learning library [11]. The dataset consists of 768 instances with 8 features which are in CSV format. In this dataset, 500 instances belong to non-diabetic class remaining 268 instances were diabetic patients. The features include pregnant count, Plasma glucose concentration, Diastolic blood pressure (mm Hg), skin thickness (mm), serum insulin (mu U/ml), body mass index, Diabetes pedigree function, and Age (years).

In our study, we used all the available features in all experiments. The data set was divided into training (60%) and testing (40%) parts. Furthermore, the training data set was further divided into 10% validation set to evaluate the performance of the model. These were randomly selected and experiments were repeated ten times to make sure there is no bias in the system. The final accuracy was calculated as the average accuracy obtained from these experiments.

III. Methodology

A. Support Vector Machine

The SVMs have proven to be very effective for various data classification tasks. It tries to find the optimal separating hyperplane between classes by finding the set of points that lie on the edge of the class descriptors. The distance between the classes is referred to as margin. SVM algorithms finds a margin such that its distance is maximum. The higher the margin, the better the classification accuracy can be obtained for the classifier. The data points lying on the border are known as support vectors (Fig. 1). Hence the name support vector machine. The rest of the training samples are discarded. SVMs can achieve good performance even on small training samples as less training samples are effectively used.

Primarily, SVMs are designed to deal with linearly separable binary classification data. Several variations have been proposed to adopt it for multi-class classification problems. Similarly, it can also be applied for classification of nonlinear cases by applying kernels techniques [12]. These kernels apply mapping from nonlinear to a linear space where it is believed that the data could be easily separated. The most commonly used kernel functions include radial basis function (RBF), polynomial, sigmoid etc. For our analysis, RBF kernel is used due to it popularity. The RBF kernel is calculated using following formula

$$K(x_i + x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$$

Where γ is gamma which is the RBF kernel's learnable parameter.

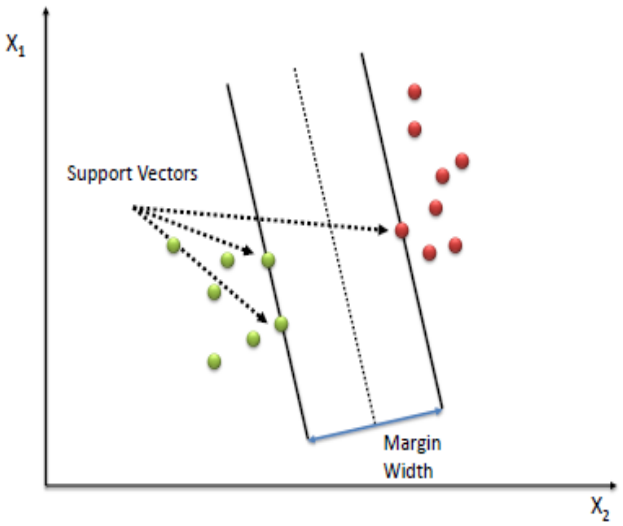

Fig. 1. Support vectors in SVM

B. *Random Forest*

The Random Forest algorithm is a supervised classification algorithm widely used in different classification tasks. This algorithm was proposed by Leo Breiman and Adèle Cutler in 2001 [13]. The random forest algorithm is derived from the decision tree classifier and, as the name suggests, is based on a set of trees where each tree depends on a set of random variables [14]. The main idea of this algorithm is explained in the Fig. 2.

The random forest algorithm is based on a set of decision trees. These different trees are characterized by the same number of nodes, but different data. The decisions of these different decision trees will be combined to give a final answer that represents an average response of all these decision trees.

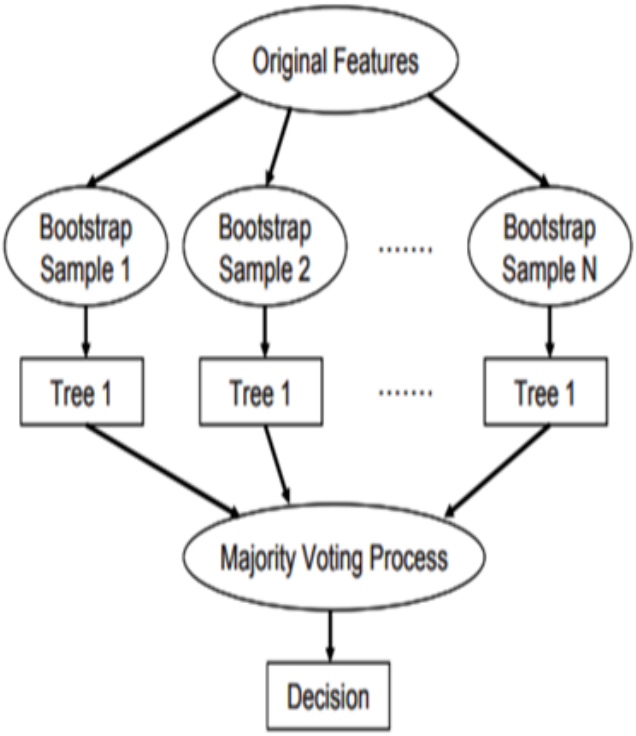

Fig. 2. Random forest algorithm [14]

C. *Convolutional Neural Network (CNN*

The CNN is one of the most commonly used DL algorithms. It is a specific type of artificial neural network that uses several layers of perceptron connected in sequence [15]. CNNs perform a series of operations on the input and transform it to produce the desired output. This output from previous layers can be taken as input to the next block.

CNNs basically consists of three main types of layers, namely convolutional layer, pooling layer and fully connected layer (fig- 3). Convolutional layer forms the core part of the network, which has local connections and weights of shared characteristics. The objective is to learn feature representations of the inputs data. The input feature maps are first convolved with a kernel and then the obtained results are passed into a nonlinear activation function. The pooling layer can be considered as a fuzzy filter, it reduces the feature dimensionality and increases their robustness. Finally, the fully connected layer takes input from previous layers and sends the signals to each neuron in it. The classification is then performed by the output layer which usually consists of a softmax classifier. For more details about the CNN, refer to [1],[11].

Like neural networks, CNNs are based on a set of neurons, weights of each neuron and biases. Each neuron receives inputs and generates an output by using an activation function. Therefore, the convolutional neuron network is a subclass of neural networks that have at least one convolutional layer. The main purpose of CNN is to reduce the network complexity that existed in Neural Networks algorithm by applying the convolution.

IV. EXPERIMENTAL RESULTS

For all experiments we selected 60% (460 samples) of the data for training and validation while 40% (154 samples) was used for testing. The performance of the proposed method was evaluated in terms of overall accuracy (OA), Kappa Coefficient (KC), precision (P), recall (R), and f-measure (F). In order to avoid any biasedness of the models, we repeated the experiment ten times and the accuracy was calculated to be the average of the all experiments. Furthermore, as the original data consisted of only 8 features, therefore, no further analysis was performed for feature selection or calculating the feature importance. We assumed all features are equally important.

Since, the classifiers have some learnable parameters, therefore, the first step that we performed was to fine tune those parameters first. Following section describes the details about the experiments and quantitative results obtained from each classifier.

A. *Support Vector Machine*

The SVM model requires tuning of few parameters. We selected radial basis function (RBF) kernel as it has proved to be effective for classification. We empirically calculated the values for gamma (1.0E-4) and C (10.0). The SVM classification produced and over all accuracy 73.94 %, precision: 62.56%, recall: 45.82%, F-measure: 51.93% and KC. The classifier produced less recall and there was higher

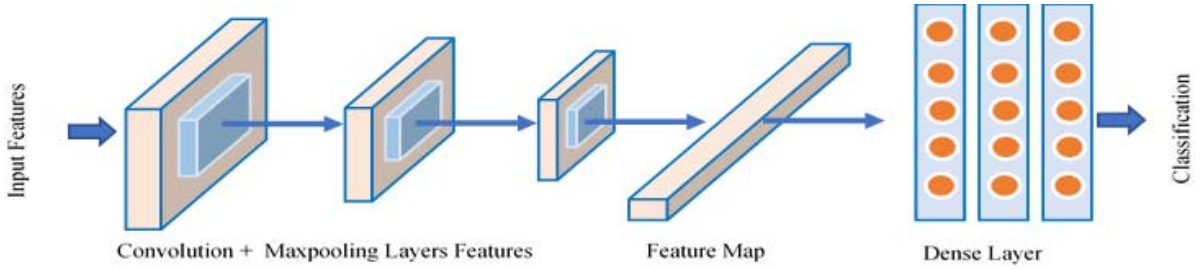

Fig 3 Architecture of the proposed CNN

error as it classified non-diabetic patients into diabetic class. Table 1 summarizes the classification results obtained for SVM classifier.

TABLE 1. CLASSIFICATION ACCURACY FOR SVM CLASSIFIER

	Non-Diabetic	**Diabetic**	**Precision**
Non-Diabetic	98	28	77.78%
Diabetic	15	24	61.54%
Class Recall	86.73%	46.15%	

B. Classification with RF

The two important parameters that affect the accuracy of RF classifier are: number of decision trees and the maximum depth. These two parameters were obtained empirically and set to 20 and 7 respectively. Fig. 4 shows the grid search ranges for each parameter and their corresponding accuracies. The overall accuracy obtained for this classier was 79.26%, precision: 84.36%, recall: 62.74%, f-measure: 70.93% and kappa: 0.556. A higher number of diabetic patients were misclassified as non-diabetic (35) out of 99 diabetic patients. However, the class precision for non-diabetic patients was relatively higher (88.43%). Table 2 shows the results obtained for RF classifier.

TABLE 2. CLASSIFICATION ACCURACY FOR RF CLASSIFIER

	Non-Diabetic	Diabetic	Class precision
Non-Diabetic	107	14	88.43%
Diabetic	35	64	64.65%
Class Recall	75.35%	82.05%	

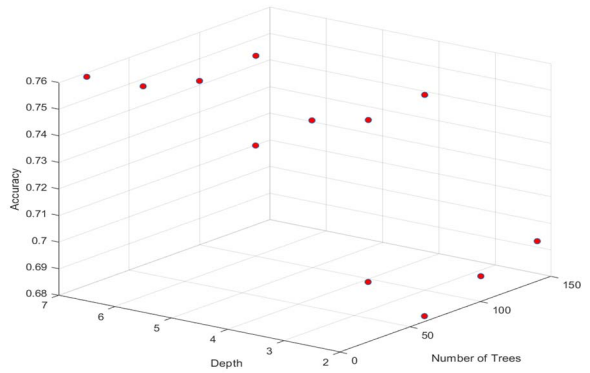

Fig. 4 Visualization of grid search for RF parameters

V. CONCLUSION

This study performed a comparative analysis of machine learning and deep learning-based algorithms for prediction of diabetes. The results showed that RF was more effective for classification of the diabetes in all rounds of experiments which produced overall accuracy for diabetic prediction to be 83.67%. The prediction accuracy for SVM reached 65.38% while DL method produced 76.81% on our dataset. In future we would like to improve the feature extraction step by applying an automatic deep feature extraction approach and for obtaining a better fitting model to improve the prediction accuracy.

REFERENCES

[1] Z. Punthakee, R. Goldenberg, and P. Katz, "Definition, Classification and Diagnosis of Diabetes, Prediabetes and Metabolic Syndrome," Can. J. Diabetes, vol. 42, pp. S10–S15, 2018.

[2] M. N. Piero, "Diabetes mellitus – a devastating metabolic disorder," Asian J. Biomed. Pharm. Sci., vol. 4, no. 40, pp. 1–7, 2015.

[3] G. Swapna, R. Vinayakumar, and K. P. Soman, "Diabetes detection using deep learning algorithms," ICT Express, vol. 4, no. 4, pp. 243–246, 2018.

[4] L. Lucaccioni and L. Iughetti, "Issues in Diagnosis and Treatment of Type 1 Diabetes Mellitus in Childhood," J. Diabetes Mellit., vol. 06, no. 02, pp. 175–183, 2016.

[5] "Type 2 Diabetes: a Review of Current Trends -," Int. J. Curr. Res. Rev., vol. 7, no. 18, pp. 61–66, 2015.

[6] K. Kannadasan, D. R. Edla, and V. Kuppili, "Type 2 diabetes data classi fi cation using stacked autoencoders in deep neural networks," Clin. Epidemiol. Glob. Heal., no. December, pp. 2–7, 2018.

[7] T. N. Joshi and P. P. M. Chawan, "Diabetes Prediction Using Machine Learning Techniques," Ijera, vol. 8, no. 1, pp. 9–13, 2018.

[8] M. T. P. Kamble, "Diabetes Detection using Deep Learning Approach," vol. 2, no. 12, pp. 342–349, 2016.

[9] S. Ramesh, R. D. Caytiles, and N. C. S. N. Iyengar, "A Deep Learning Approach to Identify Diabetes," vol. 145, no. Ngcit, pp. 44–49, 2017.

[10] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," Front. Genet., vol. 9, no. November, pp. 1–10, 2018.

[11] U. M. L. Repository, "https://archive.ics.uci.edu/ml/index.php." .

[12] N. Cristianini, J. Shawe-Taylor, and others, An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, 2000.

[13] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random Forests," no. February 2014, 2011.

[14] B. Yang, X. Di, and T. Han, "Random forests classifier for machine fault diagnosis Random forests classifier for machine fault diagnosis," no. April, 2014.

[15] S. Albawi and T. A. Mohammed, "Understanding of a Convolutional Neural Network," no. April, 2018.

[16] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Dec. 2014.