DOI 10.29042/2017-2029-2033

Helix Vol. 7(5): 2029-2033

Comparative Analysis of Pure and Hybrid Machine Learning Algorithms for Risk Prediction of Diabetes Mellitus

[1] Deeksha Kaul, [2] Harika Raju, [3] B.K. Tripathy
School of Computer Science and Engineering, VIT University, Vellore-632014, Tamil Nadu, India
[1]deekshakaul1@gmail.com, [2]harikaraju.g@gmail.com, [3]tripathybk@vit.ac.in

Received:9th June 2017, Accepted:25th June 2017, Published: 1st September 2017

Abstract
Diabetes, a chronic disease, occurs due to abnormal levels of glucose and insulin in our bodies. The major factors leading to the disease are lifestyle factors such as diet, insufficient physical activity, increased stress levels and obesity. One of the major issue with diabetes is the mildness of its symptoms which delays its diagnosis until the disease has significantly progressed leading to other medical complications like blindness and nerve and kidney damage. This paper compares the efficiency of various machine learning algorithms for predicting the onset of diabetes. The performance analysis is demonstrated on Support Vector Machines (SVM), Deep Neural Networks and Hybrid Deep Learning. The study shows that the classification accuracy is significantly higher for hybrid deep neural networks indicating its better performance. The results also justify that hybrid deep learning has better processing capability and produces accurate results, thus helping in the prediction of many diseases.

Keywords: Deep Neural Networks, Support Vector Machine, Hybrid Deep Learning, Diabetes mellitus.

Introduction

Deep learning is an unsupervised learning algorithm which mimics the human brain. The deep network has the capacity to process large datasets, implement complex functions as well as it can work with unlabeled data with small human inputs (training set) and generates the convoluted representation of unprocessed data. For enhanced performance, it has been suggested to combine supervised learning algorithm (SVM) with the deep neural networks. The unsupervised nature combined with complex learning algorithms ensures fast processing and accurate predictions for multi-dimensional data.

Diabetes is a chronic disease which in general does not have any cure except for very few rare cases. According to World Health Organization (WHO), the number of people suffering from diabetes is very high and is expected to increase due to the high prevalence of obesity and sedentary lifestyle. Diabetes leads to quite grave medical complications and sometimes death also. Diabetes occurs due to inefficient use or improper production of insulin by the pancreas. There are majorly three types of diabetes. Type-1 or juvenile-onset diabetes occurs due to insulin deficiency which is caused when the immune system attacks cells responsible for the production of insulin. Type-II or adult-onset diabetes is usually characterized by resistance to insulin. Gestational diabetes occurs during pregnancy due to high insulin levels which may or may not improve after delivery.

Diagnosis of diabetes and its type based on various symptoms is often time-consuming. Hence, there is a need for some automated process which will provide help in detecting the presence and progression of diabetes. And moreover predicting the onset of this disease by studying and analyzing the huge amounts of medical data available by applying various machine learning algorithms will immensely help the medical society and also assist in eradicating challenges faced by medical researchers. In this paper we use a hybrid deep learning model based on deep learning and support vector machine that predicts the presence Diabetes Mellitus. The model is trained using SVM followed by Deep Learning. We are effectively replacing the first layer of the deep learning architecture with a linear SVM layer. Further, this article also discusses the progression of the disease based on the standards set by the American Diabetes Association.

Literature Survey

On-going researches on deep learning techniques for machine learning, the challenges involved and their applications are discussed in [1]. It is observed by him that the challenges involved are due to the continuously evolving nature of data leading to Big Data. Wang et al [2] discuss the application of various deep learning algorithms for video analytics for object detection and tracking, face recognition and image classification in a smart city. The use of deep learning models enhances the performance of above-mentioned techniques almost to the extent of human capability due to huge amount of training data and the advanced hardware available, which have reduced training time and increased efficiency of decision making. For these reasons, deep learning will have unprecedented effects in wide range of applications in foreseeable future. Shanker and Du [3] provide an overview of application of deep learning for natural language processing. According to this research, deep learning algorithms may provide varying efficiency for different tasks. Deep learning has been applied to NLP with some success

though not very satisfactorily and shows room for improvement.

According to Wang [4], neural network and deep learning algorithms can be used to identify network traffic based on various features. Deep learning approach has shown efficient results in feature learning and protocol identification and detection. Han et al [5] have developed data mining algorithms using Rapid Miner to predict Diabetes among the population. The Pima Indians Diabetes Data Set, which records the presence of diabetes in the female population, is used to train the models. For the prediction, the authors use Decision tree and ID3 models which have 72% and 80% of accuracy respectively.

Songthung and Sripanidkulchai [6] analyze the data collected from 12 hospitals in Thailand focusing on females older than 15 years to predict the high risk of Type-2 Diabetes Mellitus. The authors use Naïve Bayes and Decision Tree classifiers for risk prediction and use coverage i.e. ability to identify individuals who will eventually be diagnosed with Diabetes, as the metric for the efficiency of the model. Kamble and Patil [7] compare the efficiency of deep learning based restricted Boltzmann machine and decision tree algorithms for identifying if a patient is diabetic. Their study concludes decision tree to be a better prediction model with lower error rates. The dataset used in their research has over 300 data points and their elaborate use of data has depicted an improvement in the accuracy of the model proposed by them.

Methodology

In this section, we present the processes of data acquisition, data pre-processing, data training and validation along with performance metrics used in our article.

3.1. Data acquisition

The data set used for experimenting and evaluation has been obtained from UCI machine learning repository. The data set is better known as Pima Indians Diabetes Dataset which was provided by National Institute of Diabetes and Digestive and Kidney Diseases and has been briefly described in Table 1.

3.2. Data pre-processing

The data present in the datasets cannot be always used directly for analysis. So the first and foremost step to be done in any data analysis process is data pre-processing which also known as data is cleansing. Data cleansing is used to remove all the errors, outliers and missing values from the data set. The dataset taken by us has null values under the attributes plasma glucose concentration, blood pressure, skinfold thickness, serum-insulin and body mass index. Instead of eliminating the tuples containing these null values, which may lead to huge loss of information, we replaced these values by the mean values of the corresponding attributes. It is observed that taking median values instead of mean values the accuracy decreases by 2%.

Various attributes have different numerical ranges which might have led to the undue influence of these attributes as they tend to have varied ranges. Thus data normalization is done to bring the values of all the attributes in the range of 0 and 1. The normalization is done as follows:

$$\text{Normalized } X = \frac{x - \min(x)}{\left(\max(x) - \min(x)\right)} \quad (1)$$

Table 1. **Attributes present in the dataset**

S. NO	NAME OF ATTRIBUTE
1	No. of Pregnancies
2	Plasma Glucose Concentration
3	Diastolic Blood Pressure (mm Hg)
4	Triceps Skin Fold Thickness (mm)
5	2- hour Serum Insulin (mu U/ml)
6	Body Mass Index
7	Diabetes Pedigree Function
8	Age
9	Class Variable (0 or 1)

To prevent over-fitting of data the dataset has been divided into training set, cross-validation set and test set in the ratio of 6:2:2.

3.3. Training and validation

After the dataset has been divided into training and test sets, in the next step we apply various machine learning algorithms. The three algorithms we used for training and validation are:

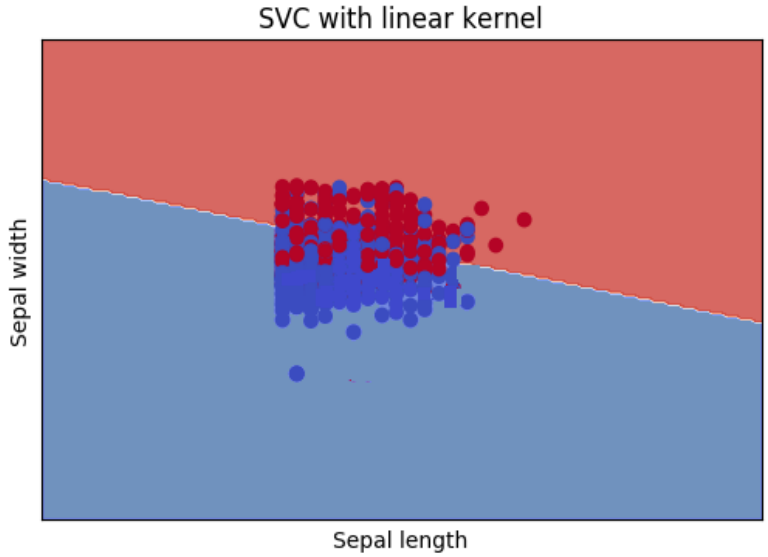
- Support Vector Machines
- Deep Neural Networks
- Hybrid Deep Learning

Support Vector Machines (SVM)

SVM is a machine learning algorithm which uses supervised learning approach and separates the data set into two clusters by determining a hyperplane such that the distances of the nearest points from the plane is maximum. Figure 1 illustrates the working principle of this algorithm. However, for SVM to be applicable, it is not necessary that the points should be linearly separable as using the kernel-trick it can produce a hyperplane for inseparable dataset also. The expression (2) is used to produce a decision boundary such that the data set is classified into positive and negative classes relative to the hyperplane $y = w.x + b$, with constraint function $y_i(x_i w + b) \geq 1$. The Lagrangian function for SVM is described as follows. Here α_i is the support value.

$$L = \frac{1}{2}\|w\|^2 - \sum_i \alpha_i \left(y_i \left(w_i x + b \right) - 1 \right) \quad (2)$$

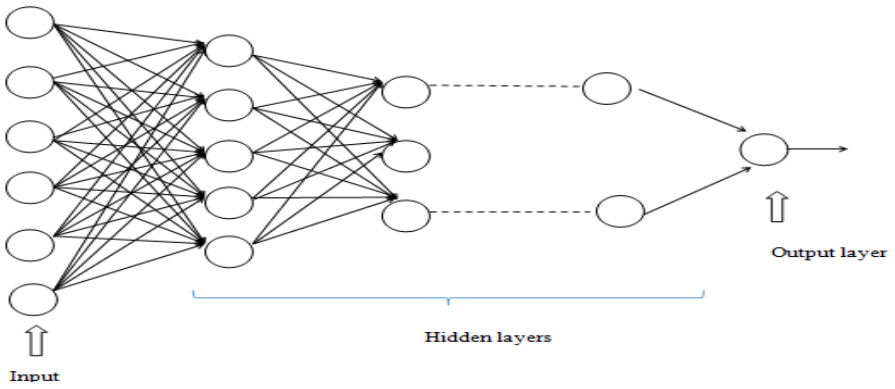
Fig 1. Support Vector Machine



Deep Neural Networks

Deep Learning is a sophisticated machine learning algorithm which has enabled several practical application of artificial intelligence. Its concept is inspired by the linking and functionalities of neurons present in the human brain. It is a multi-layer feed-forward artificial neural network which supports unsupervised learning and is trained with stochastic gradient descent using back-propagation. It has multiple layers of hidden units which make it different from a neural net with only a single hidden layer as depicted in figure 2.

Fig 2. Deep Neural Net Architecture



Hybrid Deep Learning

A hybrid model typically involves a combination of multiple clustering or classification machine learning algorithms. The discussed hybrid neural network employs deep neural networks along with support vector machines. In the first stage, the data is fed to SVM for classification and its output is used for the construction of prediction model based on deep learning. This model outperforms the contributing machine learning algorithms exhibiting better accuracy, precision, and area under the Receiver Operating Characteristics (ROC) curve and justifying its superior performance to pure SVM and pure Deep Neural Networks.

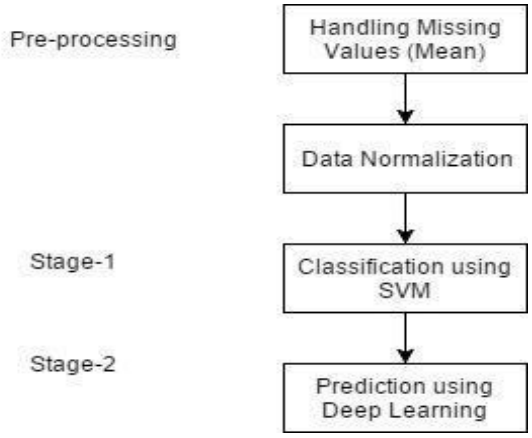


Fig 3. Framework of the proposed hybrid system

3.4 Performance metrics

We have evaluated the performances of the above-mentioned algorithms using performance metrics accuracy, precision, Root mean square error and area under ROC curve. Accuracy is defined as the proportion of a number of correct assessments to the total number of assessments. The precision is the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative. Mean Square Error (MSE) indicates the error in prediction. Lesser the MSE value, better the model. The area under the ROC curve is also an efficient performance metric. Higher the AUC value, better the performance.

Results and discussions

The initial data had several missing attribute values. We make use of average imputation to handle the missing data. After data imputation, the data is normalized and fed to various algorithms to check for efficiency. This paper uses SVM, Deep Learning and then compares the performance of these models with a hybrid model of SVM and Deep Learning. The efficiency metrics of the above-mentioned algorithms is recorded in Tables 2 and 3. Table 3 records the correctly and incorrectly predicted negative and positive values which give an idea of the accuracy of the model used. For analysis, we make use of sklearn, keras and matplotlib python libraries.

Table 2. Table of confusion

Algorithm	True positive	False positive	True negative	False negative
Deep Learning	150	50	450	118
SVM	140	46	454	1 28
Hybrid Deep Learning	187	70	430	81

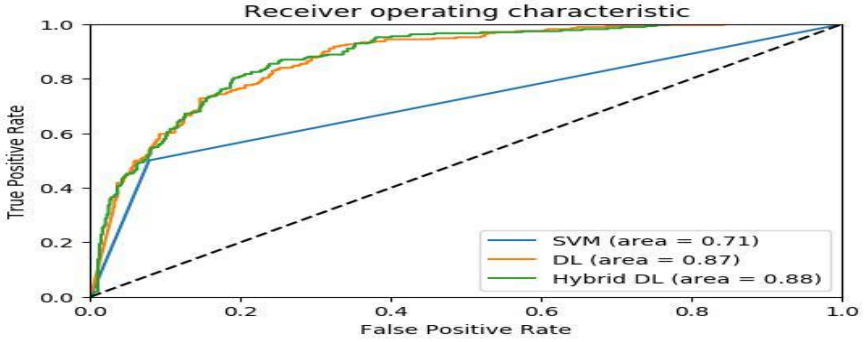


Fig 4. Comparing ROCs

As observed from Table 3, deep learning and SVM have comparable performances with a marginal difference in accuracy and precision. But with a significant difference in MSE (0.1489 and 0.2266 respectively) and AUC (0.8538 and 0.7512 respectively) values deep learning shows better performance. Comparing the performance of deep learning and hybrid deep learning, the hybrid model shows better accuracy (0.8034), AUC (0.8733) and precision (0.7615) and least MSE (0.1376). Considering the ROC curve in Fig 4, the behavior for deep learning and hybrid deep learning is similar though the area under the curve is reasonably better for hybrid model indicating its efficiency to correctly predict positive and negative classes.

Table 3. Performance analysis

Algorithm	Deep Neural Network	SVM	Hybrid Deep Neural Network
Accuracy	0.7812	0.7734	0.8034
AUC	0.8538	0.7152	0.8733
MSE	0.1489	0.2266	0.1376
Precision	0.7502	0.7526	0.7615

Table 4. Glucose Concentration and diabetes stages

Concentration	Stage
4.5-5	Stage 1
5.1-6.5	Stage 2
6.6-16	Stage 3
16.1-22	Stage 4
>22	Stage 5

Along with the diagnosis we also predict the progression of the disease according to the guidelines set by American Diabetes Association (ADA). ADA uses plasma glucose concentration to identify the level of severity of Diabetes. With increasing level of malignity, the disease can be classified into five stages based on the amount of plasma glucose present in the bloodstream and the insulin level. Table 4 describes the ranges of glucose concentration set by American Diabetes Association to understand the progression of Diabetes.

Conclusion

It can be concluded from the study that hybrid deep learning provides the most satisfactory results for prediction of diabetes. Least error rate and highest area under the ROC curve, accuracy and precision values provide evidence of better performance as compared to pure SVM and pure Deep Learning models. But even with extensive training, there is a chance of misdiagnosis as the model is not 100% accurate. For future work, the performance of the decision support system can be optimized by training with more extensive data. Newer data cleansing techniques like using Regression Substitution for handling missing values can be used for better accuracy. Furthermore, the current data only focuses on the female population and we intend to explore the trends in the male population of the society as well.

References

1. Chen, X. W., & Lin, X.: Big data deep learning: challenges and perspectives, pp. 514— 525. *IEEE Access* (2015)
2. Wang, L., & Sng, D.: Deep Learning Algorithms with Applications to Video Analytics for A Smart City: A Survey (2015)
3. Du, T., & Shanker, V. K.: Deep Learning for Natural Language Processing
4. Wang, Z.: The Applications of Deep Learning on Traffic Identification. *BlackHat USA* (2015)
5. Han, Jianchao, Juan C. Rodriguez, and Mohsen Beheshti.: Diabetes data analysis and prediction model discovery using rapidminer. In: Second International Conference on Future Generation Communication and Networking, FGCN'08. vol. 3. IEEE(2008)
6. Songthung, P., & Sripanidkulchai, K.: Improving type 2 diabetes mellitus risk prediction using classification. In: 13th International Joint Conference, pp. 1—6. Computer Science and Software Engineering (JCSSE), IEEE(2016)
7. Kamble, M. T. P., & Patil, S. T.:Diabetes Detection using Deep Learning Approach. IJIRST, vol. 2, Issue 12 (2016)
8. UCI Machine Learning Repository, https://archive.ics.uci.edu
9. Gordon C. Weir & Susan Bonner-Weir: Five Stages of Evolving Beta-Cell Dysfunction During Progression to Diabetes (2004).