Automated Detection and Classifying Diabetes Mellitus using CNN

Aswiga R.V (✉️ aswiga91@gmail.com )
Vellore Institute of Technology - Chennai Campus

KARPAGAM M
SRM Institute of Science and Technology

Chandralekha M
Amrita Vishwa Vidyapeetham Amrita School of Engineering Chennai

Sathiya Kumar C
VIT University

Deena S
Amrita Vishwa Vidyapeetham Amrita School of Engineering Chennai

Research Article

Keywords: Convolution Neural Network, Binary Classification, Machine Learning, Diabetes Mellitus

Posted Date: January 18th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2405338/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published at Soft Computing on April 11th, 2023. See the published version at https://doi.org/10.1007/s00500-023-08122-y.
Abstract

Machine learning algorithms have recently attracted more attention in the medical field and have been successfully used in a variety of medical applications. Machine learning is used to assist medical practitioners to solve complicated problems. Thus to function autonomously without seeking any assistance or help from a person, a well-known branch of machine learning called deep learning completes several related tasks parallely. It is successfully used in a variety of applications, including disease prediction and disease progression. Since tasks are assumed to be related to one another, existing learning methods adapts the Convolution neural network (CNN) to classify the diabetes mellitus disease effectively. To detect the abnormality, the proposed work uses a deep learning network that includes a novel CNN model that splits the data into separate training and testing sets before doing classification. Using CNN, the proposed work achieves better accuracy and outperforms well to the best of our knowledge.

Introduction

Machine Learning guarantees to offer solutions for lots of actual real-world medical problems. This field is drastically explored in lots of application including sickness identification, ailment diagnosis, personalized treatment, behavioral changes, and discovery of drugs and in manufacturing drugs. But the advanced deep learning plays a vital role in medical research, used to take predictions with respect to radiology and radiotherapy, gives anticipated results for being fit, and for epidemic outbreak prediction etc. Multitask learning is a branch of machine learning that focuses research in such a way that additionally allows medical examiners to research the specific information which might also additionally result in improvised diagnoses and treatment. This machine learning approaches both traditional as well as advanced deep learning techniques in this proposed work are trying to identify and classify the diabetes mellitus effectively.

Diabetes mellitus, generally referred to as diabetes, is a disorder that influences a substantial majority of human beings globally. Though this is considered as normal disease but when not controlled and monitored properly, this turns to be a life threatening disease that leads to heart attack. Diabetes cannot be cured, but can be controlled using several stages like pre-diabetes, post-diabetes and normal; Because of the living standards in the current world, diabetes is seen increasingly common in people's every day's life. Using any machine learning techniques, quickly and accurately diagnosing diabetes is a much demanded research and also a hot topic worthy studying. Normally in medicine, the diagnosis of diabetes mellitus is done according to several features like Urinecreatinine, Alb/Crea Ratio, Lipoprotein A, BUN, Apolipoprotein-B, Apolioprotein A1, Microalbumin, Serum Creatinine etc.

This disease as already mentioned earlier can be effectively treated if diagnosed properly. Machine learning and advanced deep learning technologies can help a patient and medical practitioner to track the health status and predominantly judge levels of diabetes before and after fasting according to their daily physical examination data. This will be helping doctors as this serves as a reference test values for
doctors. For advanced machine learning method, selecting the valid features and the appropriate classifiers are the hot important problem in this field. Though there are several features available, identifying diabetes using text features is considered predominantly for accurate classification. Because these text features provide sufficient information about the patient characteristics that caused the disease like age, heredity, glucose level as well as these characteristics will definitely vary from one person to other.

**Related Work**

There has been several booming results in the field of advanced deep learning and multitask learning for predicting diabetes. In the recent years, machine learning traditional models are very much popular to solve several problems like classifying images [1], processing text [2], diagnosing fault in real world dataset [3] and in applications including healthcare [4–5]. It is widespread to use both traditional and advanced ML algorithms to address the progress of disease and for disease prediction [6–8]. In the existing work, we shortly elaborate and summarize the existing work in the medical domain.

Fuzzy based ontology model is proposed by El-Sappagh et al. [9] for the treatment of Diabetes Mellitus for which a case-based justification paradigm is used. A fuzzy semantic retrieval algorithm was proposed by the authors and this fuzzy based case ontology algorithm will manage several functionalities in different forms. There are total of 60 possible diabetic instances that are used in the fuzzy ontology. Their proposed work will respond and give good classification results even for more complicated questions that are linked to the medical principles and will effectively handle ambiguous terms.

Rahim et al. [10] proposed similar fuzzy system for detecting diabetes. Fuzzy filtering and histogram is used in the preprocessing of the dataset and fuzzy edge detection algorithm is very useful in detection of retinopathy images. K-Nearest Neighbor (KNN) and Decision Tree (DT) algorithms are used for the classifying retinopathy images with accuracy of 74.6%.

The work proposed by VijiyaKumar et al. [11] is based on the random forest algorithm that helps patient by predicting diabetes with a higher accuracy. This system works on traditional approach with huge dataset hence the model is capable of instantly predicting the diabetes disease. Three different supervised machine learning techniques like Artificial Neural Network (ANN), Support Vector Machine (SVM), and Logistic Regression are used in the work done by Tejas N. Joshi et al. [12] for detection of diabetes disease at the earliest. An ensemble based supervised learning approach is introduced by Nonso Nnamoko et al. [13]. Along with this, several other classifiers are employed and their outputs are aggregated. It is shown that using this existing method, prediction of diabetes is done with highest accuracy.

Feature Selection has immense effect on efficiency of machine learning algorithm and it is considered as key factor that impacts accuracy. The text features that are used to train the machine learning model have a huge impact on results. The first and foremost significant step in the design of the machine learning model is feature selection. Feature selection have several benefits like (1) avoiding the model to
suffer from overfitting. Overfitting will occur with less repetitive data therefore this can be avoided by feature selection. (2) The second benefit is increased accuracy. (3) Using advanced deep learning techniques with fewer efficient features will defenetly decreases processing time and therefore training of the algorithms becomes easier and faster with better results. The proposed work also surveys different research works based on important feature selection and common feature identification in the study described below.

Ben David & et.al [14] have discovered an effective mechanism to correctly estimate and find the exact relationship between the tasks. Initially, N numbers of tasks are available and each task contains several common features that are shared among them. That is, same features are used to classify and find the different diseases accurately. Therefore with this added advantage there is no necessity of having same quantity of samples to classify all tasks. From this we can conclude that the number of samples in each task can be varied. Machine learning traditional algorithms will not perform well when classifying intrinsic medical dataset because of the raw and inconsistent nature.

The work proposed by Ando & et.al [15] correctly classifies commonly shared parameters in all tasks. Identifying this shared parameter is possible and the result will be reliable only when the number of tasks is large. Similar to this, the work introduced in Maurer& et.al [16] demonstrates how multitask learning algorithms that are very linear in nature will only use common operator to preprocess the dataset. After preprocessing the dataset, multiple related tasks will be learned simultaneously.

The proposed framework by Ali Jalali & et.al [17] tries to reduce the number of unwanted features. Their proposed work restricts the features that are shared among all the tasks with the model that focuses on non-convex regularization and quadratic loss function. But it is important to note that many traditional existing machine learning algorithms will struggle to predict and classify the results as many irrelevant features are shared in related tasks.

From the survey done, it is clear that the prediction of diabetes disease is outperformed well in traditional machine learning algorithms like SVM, decision trees, random forest, and fuzzy based approaches with huge dataset as well as for the image and other wavelet datasets [21–23]. Therefore the proposed work tries to implement the prediction and classification of diabetes disease using advanced deep learning technique like CNN instead of using traditional algorithms for comma separated values. This deep learning CNN model will make use of selected important text features that are very limited in size for N samples.

**Material And Methods**

The proposed convolutional neural networks follows similar strategy like supervised learning technique where they receive the input dataset, detects the essential features in each of them, and then finally train a model on it. However, convolutional neural network learns the features automatically by themselves. Hence, this becomes the tedious work for extracting and describing features automatically during the training phase, due to which the classification error is minimized when the model is trained. As mentioned
in the work proposed by [18]-[20], the concept of advanced deep learning techniques have been employed for designing CNN framework for medical domain. The existing CNN model for the medical application consists of four types of layers namely, the convolutional layer, which is the first layer; its purpose is to identify the essential features received as input. Then second layer will be the pooling layer, which is often placed in-between two convolution layers. This pooling layer receives numerous feature maps and then applies the pooling operation to each of the feature maps. This will result in reducing the size of the input while preserving their most important characteristics without changing the meaning in the dataset. The third will be the ReLU correction layer that refers to the real non-linear function in the network and finally the fully-connected layer is the last layer of a neural network that receives an input text and produces a corresponding classification output. The functionalities of each and every layer are described below. Inspired by the existing work [18]-[20], the proposed work also focuses on developing a CNN model for processing diabetes dataset. The architecture diagram of proposed framework is presented below.

The functionality of the proposed CNN layers and their uses are mentioned below. The convolutional layer function is described with the help of the following equation:

\[ y_k^n = h(b_k^n + \sum w_{n-1,k} * m_x^n) \]  \hspace{1cm} (1)

where,

- \( n \) denotes the number of layers in the coding network
- \( h \) is described as the activation function
- \( m_x \) will be input feature map
- \( y_k \) is output feature map
- \( w_{n,k} \) will be the weight kernal for CNN layers
- \( * \) is used for convolutional operation
- \( b_k \) will act as bias

Then dense layer operation is defined by,

\[ \text{output} = \text{activation} \left( \text{dot} \left( \text{input, kernel} \right) + \text{bias} \right) \]  \hspace{1cm} (2)

Table 1 explains about the terminologies used in the dataset.

Table 1. Terminologies used in the dataset
<table>
<thead>
<tr>
<th>Abbreviated Term</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGOT</td>
<td>Serum glutamic oxaloacetic transaminase. This is present in heart and liver cells</td>
</tr>
<tr>
<td>SGPT</td>
<td>Serum Glutamic Pyruvic Transaminase. This is present in body and cell tissues.</td>
</tr>
<tr>
<td>GTT</td>
<td>The glucose tolerance test</td>
</tr>
<tr>
<td>TGL</td>
<td>Triglycerides. This is a type of fat present in blood.</td>
</tr>
<tr>
<td>HDL</td>
<td>High-density lipoprotein</td>
</tr>
<tr>
<td>LDL</td>
<td>Low-density-lipoprotein (LDL) cholesterol. This is bad cholesterol.</td>
</tr>
<tr>
<td>hs_CRP</td>
<td>High sensitivity C-reactive protein (Useful in type 2 diabetes prediction</td>
</tr>
<tr>
<td>SOD</td>
<td>Superoxide dismutase (SOD). This is found in food sources.</td>
</tr>
<tr>
<td>LPO</td>
<td>Lipid peroxidation (LPO)</td>
</tr>
<tr>
<td>APOB/APO</td>
<td>Apolipoprotein</td>
</tr>
<tr>
<td>TSH3</td>
<td>A thyroid-stimulating hormone. This will reflect the abnormalities in diabetes</td>
</tr>
</tbody>
</table>

**Experiment And Results**

**Data set Used**

The medical dataset which is used for the proposed work consists of 38 similar features that are shared among three tasks. The dataset consists of three tasks. Each task separately consists of 38 features. This feature set remains same for all tasks. The features used in our data set are age, Alb/Crea Ratio, SGOT, SGPT, fasting, postprandial, GTT 1 ½ hr, GTT 2 hr, GTT ½ hr, GTT 1 hr, HbA1c, MGV, Insulin, T.Cholesterol, TGL, HDL, LDL, Apolipoprotein A1, Lipoprotein A, BUN, Serum Creatinine, Alk.Phosphatase, T.Protein, Albumin, Cholesterol/HDL ratio, Apo lipoprotein-B T.Bilirubin, D.Bilirubin, Hemaglobin, TSH3, Microalbumin, Urinecreatinine, Uric Acid, Homosystiene, hs_CRP, SOD, LPO, and APOB/APO. There are totally 150 samples of patient's data set containing 38 features.

**Evaluation Metrics**

The metrics chosen for evaluating the proposed model is ‘Accuracy’. Since it is the medical dataset, the main goal is to clearly differentiate the patient identified as diabetic and normal case. Higher the accuracy, better the model in diagnosing disease. Accuracy is calculated by dividing number of correct predictions by total number of predictions. The proposed work results are shown in the graph below.

**Conclusion**
The framework proposed in this work handles medical dataset related to diabetes which is limited in number. The proposed framework handles csv values. The model will be further improved by using additional feature extraction techniques. The proposed approach achieves good diagnosis accuracy even when the dataset consists of lot of diversities by its nature. The proposed model outperforms without any misclassifications and reduces loss when comparing to the existing state of the art method. Most of the existing work is based on the traditional machine learning approaches where as the proposed work is based on deep learning network that handles intrinsic dataset effectively. As a future work, more sophisticated feature extraction algorithms can be developed to enhance the accuracy of dataset that consists of numerical measures.

**Declarations**

**Ethical approval**

This article does not contain any studies with human participants or animals performed by any of the authors. Informed consent was obtained from all individual participants included in the study.

**Funding**

This study was not funded by any funding agency

**Conflict of interest**

All authors declare that they have no conflict of interest.

**Informed Consent**

Not applicable

**Authorship Contribution**

Not applicable.

**References**


Figures

![Diagram of multistage transfer learning technique](image)

**Fig 1.**

Figure 1

Legend not included with this version
Fig 2.

Figure 2

Legend not included with this version
Fig 3.

Figure 3

Legend not included with this version