Literature Review on Diabetes Detection using Machine Learning

Insert Subtitle Here

Seif ElAnsary†  
 CSCE Department  
 The American University in Cairo  
 ID:900221511  
 seif\_elansary@aucegypt.edu

MagdElDin AbdalRaaof  
 CSCE Department  
 The American University in Cairo  
 ID: 900211145  
 MagdElDin@aucegypt.edu

FirstName Surname  
 Department Name  
 Institution/University Name  
 City State Country  
 email@email.com

INTRODUCTION

Diabetes Mellitus (DM) is one of the most prevalent diseases in the entire world. According to the World Health Organization (WHO), the number people affected by DM increased from 200 million to 830 million from 1990 to 2022 [1]. DM, if not treated correctly, can cause various health complications including but not limited to: blindness, amputations, heart attacks and kidney failure [1].

DM can take on different forms, most commonly, type 1 and type 2. Type 1 results from your body not producing insulin due to the fact that your immune system attacks the cells responsible for making insulin. Insulin is essential for our lives since it is the thing that regulates glucose in our body, the very thing that gives us energy to do our daily tasks. Type 2 results from your body not using the insulin produced properly or your pancreas can not make enough of it [2].

The deadly complications and worldwide spread of DM serves as our motivation in using machine learning techniques in order to detect diabetes in people who were not previously diagnosed and even predict if a person will develop DM or not. Treatment of DM can decrease the risk of eye disease, kidney disease, and nerve disease by 40%, and reduce the risk of heart diseases and strokes by up to 50% [3]. Treatment is a matter of life or death for many individuals with the disease, making our work of great importance.

LITERATURE REVIEW

In Ioannis Kavakiotis’s systematic review of machine learning in Diabetes research, genetic background and environmental factors were used to train common classification algorithms like Support Vector Machines (SVMs), Artificial Neural Networks (ANN) and Decision Trees (DT) to identify if a person has DM [4]. The study showed that SVMs produced the best classification results, up to 80% accuracy. However, the authors were wary of the datasets that were used in the training of these algorithms. This is due to the fact that the dimensionality and the ratio of instances to features could produce varying results that perform poorly on other datasets. Moreover, Machine learning algorithms were used in predicting a personalised treatment plan for a patient diagnosed with DM. This was partially calculated using the patients’ blood glucose levels and the amount of insulin in their blood [4].  
  
Another paper that used machine learning to diagnose DM used principal component analysis (PCA) and adaptive neuro-fuzzy inference system (ANFIS) to create an expert system that does this task [5]. PCA was used for dimensionality reduction of the features and ANFIS was used to set the weights of a Tagaki-Sugeno-Kang fuzzy rule-based system. The author, however, used the accuracy and specificity as measures for how good the system was without relying on the precision and recall, giving dubious results for the rule-based system.  
  
There was a paper that compared various machine learning techniques in detecting diabetes in patients. It used AdaBoost, logistic regression, KNN and a Perceptron and various other techniques [6]. However, the paper only used a dataset containing 800 rows and 10 features, which is highly likely to produce a subpar classifier. This was illustrated when the authors used an external dataset to test their classifiers. Even though, on average, the models scored around 77%, in regards to the importance of correctly diagnosing patients, this accuracy is definitely not good enough to be deployed to real life situation.

The literature on the topic seems to be circling around the same methodology. In paper [7], they used Logistic Regression and Gradient Boosting Machine (GBM) on a record containing approximately 880,000 patients in order to predict if a patient has DM. What we liked about this paper was the use of complex laboratory information taken from samples of the respective patients. They used: “(age, sex, fasting blood glucose, body mass index, high-density lipoprotein, triglycerides, blood pressure, and low-density lipoprotein)”. I believe that using more descriptive and specialised features that describe the problem and are related to it helps immensely in the training of any machine learning model. Moreover, the number of records is more than sufficient to train a good, descriptive model with high performance. That was evident in this paper, for the AUC of the GBM model was 84.7%, and for the Logistic Regression, it was 84%.

METHOD  
  
From the literature review, we have decided to use logistic regression as our base machine learning model. The dataset we have is numeric, which should make training the model considerably easier, since we will not have to do much preprocessing and no need to do one-hot encoding or any other types of encoding. We will use the dataset in 1.4 for reasons that will be explored later.

DATASET REVIEW

Multiple datasets have been considered for this project and each had their own advantages and disadvantages.

1.0 CDC’s U.S. Chronic Disease Indicators (<https://healthdata.gov/dataset/U-S-Chronic-Disease-Indicators/dhcp-wb3k/about_data>)

The first dataset was found on HealthData.gov in a publicly available manner and is a dataset of 115 indicators developed by the CDC for different chronic illnesses. The data types range from numerical, to textual, to even geolocational data [8]. There are 309,216 rows and 34 features. Examples of the features include Topic (the chronic illness in question), StratificationCategoryID1 (referring to the first category people were separated on in this row), YearStart (probably in reference to the year the particular questioning in the row started) [8].

The data set, while rich in sample size, has no clear label. Additionally, the features are only occasionally descriptive if the person reading the dataset is not aware of the context behind the dataset. Moreover, the dataset tackles more than just diabetes and the other chronic illnesses in the set can cause issues, especially since there is no clear information about individual patients.

1.1 Arjun9603’s Diabetes\_Prediction (<https://www.kaggle.com/datasets/arjun9603/diabetes-prediction>)

Our second dataset was found publicly on Kaggle with a useful CC0: Public Domain license [9]. Presumably it is a dataset of different health indicators affecting the likelihood of a person in India having diabetes. The data types are mostly numerical with only one Boolean data type. There are only 767 rows and 9 features [9]. Examples of the features include 6, 148, 72, and 35. The label is presumably the column called 1 since it is the only Boolean column.

This dataset is lacking in almost every measure we are looking at. The dataset has a very small number of rows and a lacking number of features. Furthermore, the feature names are not descriptive and thus cannot be reasonably interpreted by anyone other than the dataset creators. The dataset could have been useful if it were not for these glaring flaws.

1.2 Ankit Batra’s Diabetes Dataset (<https://www.kaggle.com/datasets/ankitbatra1210/diabetes-dataset>)

Another dataset worth noting is also found on Kaggle with a CC0: Public Domain license to make access and availability easy [10]. The data was collected through conducting surveys and collecting information from clinical studies, public health databases, hospitals, and labs [10]. The data types in the features are strings, integers, and Booleans [10]. There are around 70 thousand rows and 34 features [10]. The dataset has 34 features, all of which are highly descriptive and arguably dive deeper into the details of how a person’s medical history can contribute to risk of developing specific types of diabetes. Some of the features include Genetic Markers (an indicator of whether genetic markers linked with diabetes are found), Family History (whether the family has a known history of diabetes), Insulin Level (measured in microunits per milliliter), Age, and Digestive Enzyme Levels [10]. There are This dataset delves deeper into the various types of diabetes a person can suffer from, with the label feature called Target accounting for 13 different types of diabetes [10].

The dataset is relevant in concept and could have been useful if it were not for two main factors. The first factor that causes issues is the low number of rows at only 70 thousand rows. The second element is that the dataset features exclusively people who were diagnosed with one type of diabetes or another, so people without diabetes are not present in this dataset. Our goal is to predict whether a person has diabetes or not, so it is important to have samples without diabetes.

1.3 Mohammed Mustafa’s Diabetes Prediction Dataset (<https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data>)

The next dataset is also a Kaggle dataset with an inconvenient copyright on the dataset belonging to its authors [11]. The dataset compiles medical and demographic patient data and whether the patient has diabetes and was collected using electronic health records [11]. The data types are mainly integers with a few decimal features and a couple of string features [11]. It has around 100 thousand rows and 9 features including the label [11]. Some of the features include age, gender, hypertension (whether a person has elevated artery blood pressure), HbA1c\_level (referring to the measure of average blood sugar level over the last 2 to 3 months), and blood\_glucose\_level [11]. The label is a column called diabetes which is Boolean [11].

The use of these health records provides a more relevant view of reality considering they are often used in real life [11]. However, the small number of features combined with the limited number of rows at around 100 thousand rows only makes this dataset unusable for our purposes. Another unfortunate fact is that the data is copyrighted by its original authors despite being available online which may cause additional issues.

1.4 Alex Teboul’s Diabetes Health Indicators Dataset (<https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>)

This last dataset was found on Kaggle with a CC0: Public Domain [12]. The data was collected through the Behavioral Risk Factor Surveillance System (BRFSS) which is an annual survey the CDC conducts through the telephone to collect information regarding behaviors that can pose a risk to a person’s health, any chronic conditions the surveyed individual may suffer from, and how they use services meant to prevent health complications [12]. The data found in this dataset is all decimal or Boolean [12]. The dataset contains 3 files, the first of which contains 253,680 rows with 21 features with the label being whether the person is diabetic, nondiabetic, or prediabetic [12]. The file is not balanced as warned by the author of the dataset. The second file only contains 70,692 responses and has the same 21 features with the only differences being that the label merged diabetic and prediabetic into one value and that the file is balanced between diabetics and non-diabetics. The third file is like the first file, except for the change to the label so that diabetics and prediabetics share a value while non-diabetics have another [12]. This means the first file has 3 classes for the label while the other files seem to be built around training a binary classifier.

The columns are descriptive while being easy to understand by looking at the explanation attached to each. Examples of the features include HighBP (representing whether the blood pressure of the respondent is high or not), BMI, Smoker (whether or not the respondent smoked more than 100 cigarettes in their life), HeartDiseaseorAttack (whether or not the respondent suffers from heart disease or has had a heart attack), PhysActivity (whether the respondent has had physical activity outside of their job in the last month), and NoDocbcCost (whether or not a person needed to visit a doctor but was unable due to cost) [12]. The label here is the Diabetes\_012 column in the multi-class file and Diabetes\_binary column in the binary class file [12].

Such detailed features that cover a large variety of health and personal information gives us enough data points to consider training our model. Furthermore, it is the only dataset so far with at least 250 thousand rows and at least 15 features. A minor limitation is how many of the fields are binary in nature due to inherent limitations of a phone survey. Another minor limitation is that some of the non-binary fields require normalization, but that should not be a notable problem.

REFERENCES

[1] World Health Organization. 2024. Diabetes. *World Health Organization*. Retrieved February 17, 2025 from <https://www.who.int/news-room/fact-sheets/detail/diabetes>

[2] Diabetes UK. 2017. Types of diabetes. *Diabetes UK*. Retrieved February 17, 2025 from <https://www.diabetes.org.uk/about-diabetes/types-of-diabetes>

[3] CDC. 2024. Health and economic benefits of diabetes interventions. *National Center for Chronic Disease Prevention and Health Promotion (NCCDPHP)*. Retrieved February 17, 2025 from <https://www.cdc.gov/nccdphp/priorities/diabetes-interventions.html>

[4] Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, and Ioanna Chouvarda. 2017. Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal* 15, (2017), 104–116. DOI:https://doi.org/10.1016/j.csbj.2016.12.005

[5] Kemal Polat and Salih Güneş. 2007. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. *Digital Signal Processing* 17, 4 (July 2007), 702–710. DOI:https://doi.org/10.1016/j.dsp.2006.09.005

[6] Aishwarya Mujumdar and V Vaidehi. 2019. Diabetes Prediction using Machine Learning Algorithms. *Procedia Computer Science* 165, (2019), 292–299. DOI:https://doi.org/10.1016/j.procs.2020.01.047

[7] Hang Lai, Huaxiong Huang, Karim Keshavjee, Aziz Guergachi, and Xin Gao. 2019. Predictive models for diabetes mellitus using machine learning techniques. *BMC Endocrine Disorders* 19, 1 (October 2019). DOI:https://doi.org/10.1186/s12902-019-0436-6

[8] CDC. 2024. U.S. Chronic Disease Indicators. *HealthData.gov*. Retrieved February 19, 2025 from <https://healthdata.gov/dataset/U-S-Chronic-Disease-Indicators/dhcp-wb3k/about_data>

[9] arjun9603. 2024a. Diabetes\_prediction. (December 2024). Retrieved February 19, 2025 from <https://www.kaggle.com/datasets/arjun9603/diabetes-prediction>

[10] Ankit Batra. 2024. Diabetes dataset. (September 2024). Retrieved February 19, 2025 from <https://www.kaggle.com/datasets/ankitbatra1210/diabetes-dataset>

[11] Mohammed Mustafa. 2023. Diabetes prediction dataset. (April 2023). Retrieved February 19, 2025 from <https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data>

[12] Alex Teboul. 2021. Diabetes health indicators dataset. (November 2021). Retrieved February 19, 2025 from <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00