Diabetes Detection With Machine Learning

author-list

March 18, 2025

Abstract

Diabetes is an endocrine disease that results in presence of excess blood sugar, which leads to a multitude of complications. The past decade has seen an alarming increase in diabetes prevalence, with the number of people with diabetes expected to exceed 570 million by this year. In parallel, recent years have seen many leaps in development and performance of machine learning models. This study is an attempt at harnessing these models to predict diabetes, using a well-established benchmark dataset. Existing models suffer from being unable to differentiate between different types of diabetes due to lack of data points in a single dataset. Our model hopes to alleviate this weakness by compounding two datasets with adequate data, such that one predicts the presence of diabetes, and the other, the different types of diabetes. Prediabetes, Type 1 Diabetes, Type 2 Diabetes, and Pancreatogenic Diabetes were chosen for detection. Naive Bayes Classifier, K-Nearest Neighbours, Logistic Regression, Random Forest and XGBoost algorithms were used, with XGBoost showing the most promise. The XGBoost model was thus selected, and its ¡¿ was used to obtain a feature importance graph in order to validate the outcome against existing data, and further interpret the patterns among different diabetes types. A model like this could potentially be used for diabetes detection using readily available parameters.

1 Introduction

Diabetes mellitus, commonly known as diabetes, is an umbrella term used to refer to a group of endocrine diseases.

It is a chronic condition where the body does not produce enough insulin, or cannot effectively use the insulin it produces, leading to high blood sugar levels [1]. There are multiple factors that can cause diabetes, such as pancreatic cancer, pancreatitis, genetic defects, and surgery. Apart from these medical factors, unhealthy dietary patterns, socio-economic development, and sedentary lifestyles have been identified as determinants that are driving an increase in prevelance of diabetes [2].

It is estimated that 537 million people in the age range of 20 to 79 are affected by diabetes, and this number will grow to 643 million by 2045. A study by Kumar et al. (2021) concluded that incidence of diabetes in India would grow from 9.6% in 2021 to 10.9% in 2045 [3].

Resistance to, or an insufficient amount of insulin causes suboptimal conversion of food to energy, resulting in increased hunger levels, called as polyphagia. Further, the kidney utilises more water to filter the excessive glucose in the bloodstream causing abnormal urination frequency (polyuria) and excessive thirst (polydipsia).

The World Health Organization classifies diabetes into 6 categories. Among these are Type 1 Diabetes and Type 2 Diabetes have the highest prevelance. A Prediabetic state has been identified, as an early stage of diabetes. Further, diabetic states that occur as a consequence of pancreatic diseases have been grouped under Pancreatogenic Diabetes, also termed as Type 3c diabetes.

Type 1 Diabetes is typically a consequence of autoimmune destruction, causing the pancreatic β cells to stop producing insulin. Due to its autoimmune nature and by extension, genetic predisposition, it is likely to occur at a younger age compared to other types of diabetes. People with this type of diabetes are at a higher risk of developing other autoimmune disorders. They require external doses of insulin for survival [4].

On the other hand, Type 2 Diabetes is characterised by insulin resistance and a progressive lack of insulin. Initially, the pancreas compensates by producing more insulin, but over time, it becomes unable to keep up with the demand, causing the blood sugar to rise. Although some people are more genetically prone, it also heavily depends on lifestyle factors, like lack of exercise and obesity. It accounts for nearly 90% of all diabetes cases. Type 2 Diabetes may lead to severe complications, such as cardiovascular diseases, kidney damage, nerve damage and retinopathy.

Prediabetes is a considered a precursor to Type 2 Diabetes, where the blood sugar level is high, but not high enough to be classified into the latter. People in this stage show many of the common symptoms of diabetes like polyuria and polyphagia. A study by Schlesinger et al. showed that prediabetes markedly increased the risk for incidence of cardiovascular, renal, hepatic failurea, as well as the risk of cancer and dementia [5]

Pancreatogenic diabetes is the most common after Type 2 Diabetes. Patients with acute pancreatitis are at a 34.5% risk of developing diabetes. [6] A study by Shivaprasad et al. concluded that the mortality and morbidity is higher for Type 3c Diabetes than Type 2 Diabetes [7].

The recent rise in artificial intelligence has had an impact on the healthcare field, with analytical and machine learning models processing the available health records in order to predict patient outcomes and diagnose patients more efficiently. It has accelerated the efficiency and accuracy of diagnosis by enabling data-driven decisions with greater confidence. AI systems can analyze vast amounts of medical data—such as electronic health records (EHRs), medical imaging, and genomics, allowing for early detection of conditions like cancer, heart disease, and neurological disorders, often at stages when treatments are more effective.

However, the healthcare field poses its own set of challenges to Artificial Intelligence. Availability of credible datasets, especially in economically challenged areas is scarce. This might inadvertently lead to a model that is biased due the prevalence of data from socio-ethnic groups with higher accesibility to healthcare facilities. Data privacy is yet another concern, and the abstract nature of many machine learning models raises questions about their internal working and their reliability.

Various machine learning models have been employed for diabetes prediction. K-Nearest Neighbours (KNN) has been used due to its effectiveness in handling non-linear data, achieving moderate accuracy in binary classification tasks. Naive Bayes uses a probabilistic approach, but it often struggles with feature dependencies. XGBoost has emerged as the leading algorithm, utilizing gradient-boosted trees to achieve high accuracy and robust performance on imbalanced datasets. Random Forest is yet another important algorithm, providing interpretability through feature importance. Despite these advancements, most models focus on binary classification (diabetes vs. non-diabetes) and fail to differentiate between diabetes types, such as Type 1, Type 2, Prediabetes, and Pancreatogenic Diabetes. This limitation underscores the need for more comprehensive approaches, such as the one proposed in this study.

This limitation stems from the lack of datasets that include sufficient data points for type-specific classification. To bridge this gap, this study proposes a novel two-stage approach : An initial binary classification using the Pima Indians dataset, and a subsequent multi-class classification using a secondary dataset to distinguish between Type 1 Diabetes, Type 2 Diabetes, Prediabetes, and Pancreatogenic Diabetes. By developing the model in this manner, we address the critical limitation of existing approaches, which lack the granularity to differentiate between diabetes subtypes, thereby enabling more precise predictions.

2 Literature Review

Many attempts have been made to successfully integrate ML techniques for detection of diabetes, and several different techniques have been used for the same.

KNN predicts the class of a data point by looking at the classes of its nearest neighbours in the feature space. It is a simple, non-parametric method that relies totally on the data's structure. While it gives a straightforward interpretation, it is sensitive to the choice of distance metric and the number of neighbors. Logistic Regression predicts the probability of a binary outcome using a sigmoid function. It learns the relationship between input features and the log-odds of the target variable, making linear classifier. However, it assumes a linear relationship between features and the target, which can limit its performance on complex datasets. Decision Trees split the data into regions based on feature values, aiming to maximize the purity of each region. They are intuitive and easy to interpret, but they tend to overfit the training data, especially when grown too deep. To counter this, a Random Forest classifier can be used, that combines multiple decision trees, each trained on a random subset of the data and features. The final prediction is made by averaging or taking a majority vote from all trees. This ensemble approach makes Random Forest robust and reliable for classification tasks.

It has been shown that ensemble techniques perform better than logistic regression, due to the fact that LR sufferrs from assuming linear relationships and is much more sensitive to outliers. Rajendra and Latifi concluded that the ensemble techinques Max Voting and Stacking fare better than Logistic Regression, with the latter having an accuracy 93.04% of to the former's 74.03 when trained without feature selection [8]. Seto et al. provided evidence that Gradient Boosted Decision Trees (GBDT) and logistic regression models perform somewhat equally upto training data size of 10^4, GBDT shows a significant increase after this number, while LR the model's metrics become saturated [9].

Salem et al. utilised the basic K - nearest neighbours and two of its variants which allow partial membership in multiple classes, Fuzzy KNN nand TFKNN. TFKNN turned out to be the best performer with 94.13% specific accuracy [10]

References

[1] Gojka Roglic. Who global report on diabetes: A summary. *International Journal of Noncommunicable Diseases*, 1(1):3–8, 2016.

[2] Alfredo Caturano, Margherita D'Angelo, Andrea Mormone, Vincenzo Russo, Maria Pina Mollica, Teresa Salvatore, Raffaele Galiero, Luca Rinaldi, Erica Vetrano, Raffaele Marfella, et al. Oxidative stress in type 2 diabetes: impacts from pathogenesis to lifestyle modifications. *Current Issues in Molecular Biology*, 45(8):6651–6666, 2023.

[3] Arvind Kumar, Ruby Gangwar, Abrar Ahmad Zargar, Ranjeet Kumar, and Amit Sharma. Prevalence of diabetes in india: A review of idf diabetes atlas 10th edition. *Current diabetes reviews*, 20(1):105–114, 2024.

[4] Fatima Z Syed. Type 1 diabetes mellitus. *Annals of internal medicine*, 175(3):ITC33–ITC48, 2022.

[5] Sabrina Schlesinger, Manuela Neuenschwander, Janett Barbaresko, Alexander Lang, Haifa Maalmi, Wolfgang Rathmann, Michael Roden, and Christian Herder. Prediabetes and risk of mortality, diabetes-related complications and comorbidities: umbrella review of meta-analyses of prospective studies. *Diabetologia*, pages 1–11, 2022.

[6] Diego García-Compeán, Alan R Jiménez-Rodríguez, Juan M Muñoz-Ayala, José A González-González, Héctor J Maldonado-Garza, and Jesús Z Villarreal-Pé rez. Post-acute pancreatitis diabetes: A complication waiting for more recognition and understanding. *World Journal of Gastroenterology*, 29(28):4405, 2023.

[7] Channabasappa Shivaprasad, Yalamanchi Aiswarya, Shah Kejal, Atluri Sridevi, Biswas Anupam, Barure Ramdas, Kolla Gautham, and Premchander Aarudhra. Comparison of cgm-derived measures of glycemic variability between pancreatogenic diabetes and type 2 diabetes mellitus. *Journal of diabetes science and technology*, 15(1):134–140, 2021.

[8] Priyanka Rajendra and Shahram Latifi. Prediction of diabetes using logistic regression and ensemble techniques. *Computer Methods and Programs in Biomedicine Update*, 1:100032, 2021.

[9] Hiroe Seto, Asuka Oyama, Shuji Kitora, Hiroshi Toki, Ryohei Yamamoto, Jun'ichi Kotoku, Akihiro Haga, Maki Shinzawa, Miyae Yamakawa, Sakiko Fukui, et al. Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data. *Scientific reports*, 12(1):15889, 2022.

[10] Hanaa Salem, Mahmoud Y Shams, Omar M Elzeki, Mohamed Abd Elfattah, Jehad F. Al-Amri, and Shaima Elnazer. Fine-tuning fuzzy knn classifier based on uncertainty membership for the medical diagnosis of diabetes. *Applied Sciences*, 12(3):950, 2022.