Decoding Diabetes: Understanding Risk Factors

2107794

CS430 Foundations of Data Analytics, January 2024

This paper provides a detailed analysis of diabetes risk factors using the US Centers for Disease Control and Prevention's Behavioral Risk Factor Surveillance System (BRFSS) data. Using logistic regression, it aims to predict diabetes in patients and identify key risk factors. Significant findings include the roles of high blood pressure, high cholesterol, general health, and heart disease in predicting diabetes. The study shows the effectiveness of data-driven methods in healthcare, achieving about 75% accuracy in predicting diabetes. These results are crucial for healthcare professionals and policymakers in creating focused diabetes prevention and management strategies, underlining the significance of cardiovascular health and lifestyle changes.

I. Introduction

Diabetes is a chronic condition impacting millions across the world and presents a significant public health challenge. Characterised by the body's inability to regulate blood glucose levels, diabetes can lead to severe complications such as heart disease, vision loss, kidney disease, stroke and lower limb amputation and requires continuous care [1, 2].

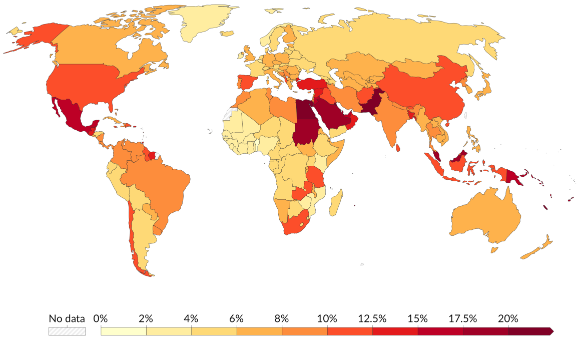

Figure 1: Diabetes prevalence around the world, 2021, the percentage of people aged 20-79 who have diabetes [6].

Figure 1 shows diabetes prevalence across the world, which casts a wide net of societal implications beyond individual health, ranging from economic burdens, societal disparities and strains on healthcare systems. Diabetes treatment consumes a significant portion of healthcare budgets, encompassing hospitalisation, medication, and ongoing monitoring. In 2022, diabetes cost an estimated $300 billion in direct medical costs in the US [3]. Diabetes also impacts working adults, leading to absenteeism, presenteeism (reduced productivity while at work), and early retirement. The economic losses due to reduced workforce participation are substantial, in 2022, indirect costs of diabetes were estimated to be about $100 billion in the US [3]. Increased healthcare needs and lost income can also burden social safety nets like disability programs and the US Medicaid system. Disparities in access to quality healthcare and diabetes education can exacerbate existing inequalities in health outcomes for vulnerable populations [4]. The growing diabetes prevalence also strains healthcare systems, pushing healthcare professionals to their limits and potentially compromising the quality of care. It shifts the focus of healthcare systems towards managing chronic conditions rather than prevention, posing challenges in resource allocation and service delivery [5].

The Centres for Disease Control and Prevention (CDC) reports that as of 2018, 34.2 million Americans have been diagnosed with diabetes, and an alarming 88 million are living with prediabetes, often unaware of their condition [7]. This study aims to address this critical health issue by leveraging data from the CDC's Behavioural Risk Factor Surveillance System (BRFSS), the nation's system of health-related telephone surveys. The BRFSS, conducted annually by the CDC, gathers data from over 400,000 Americans, across all 50 states of the country, as well as the District of Columbia and three U.S. territories, offering insights into health-related risk behaviours, chronic health conditions, and preventive service usage [8].

The individual health problems and the societal implications caused by diabetes necessitate a multifaceted approach. Beyond individual lifestyle choices, these problems must be tackled systematically [9]. Promoting healthy lifestyles, improving access to preventive care, and addressing social determinants of health are crucial for diabetes prevention and control [10]. This requires well-informed and intelligent policy-making in the relevant administrative bodies of a country and identification of the most impactful risk factors leading to diabetes, so that these may systematically be attacked and addressed. Optimising healthcare delivery, ensuring equitable access to quality care, and supporting healthcare professionals are also essential to effectively manage the diabetes burden, the BRFSS survey provides a way to detect patients presenting high risk of diabetes and may facilitate better early detection and prevention. Patient data may reveal core risk factors that can help streamline surveys like the BRFSS to focus on diabetes with more efficient and quicker questionnaires, to allow for a more comprehensive understanding of the country's diabetic population and those at risk.

Concretely, this study seeks to address these questions:

1. Can BRFSS survey data accurately predict an individual's diabetes status?
2. What are the most predictive risk factors for diabetes?
3. Can feature selection yield a streamlined set of BRFSS questions that accurately identify high diabetes risk?

With diabetes-related costs approaching 400 billion dollars annually in the US and the disease's prevalence intertwined with social determinants of health, the urgency for such a study is clear. This analysis aims not only to identify key risk factors but also to offer insights into preventive measures. Early detection and lifestyle interventions are critical in mitigating the disease's impact [11], making the development of predictive models a crucial step in guiding public health strategies and individual healthcare decisions.

II. The Dataset

A. Dataset Overview

In this study, data from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) was used. The CDC, as the US's national public health agency and part of the Department of Health and Human Services, provides reliable data. It aims to protect public health by controlling and preventing disease, injury, and disability. The BRFSS survey collects data on behaviours and health factors affecting chronic diseases and preventable conditions in adults, including diabetes. Data was gathered through phone interviews with a random sample of adults. In 2015, about 441,455 people responded to various health questions. For this study, a refined portion of this 2015 data, featuring responses from 70,692 individuals and available on Kaggle, was used [12]. This subset focuses on diabetes and its risk factors, with an equal representation of individuals with and without diabetes. This 50-50 split in the target variable helps minimise bias, keeps accuracy reliable as a metric, and reduces the risk of overfitting. The dataset includes men and women aged 18 to over 80 from all 50 US states, covering 17 factors and a binary target variable for diabetes status (0 for no diabetes, 1 for prediabetes or diabetes).

As a telephone survey relying on individuals' subjective judgement, this data might not precisely represent health factors and could introduce self-reporting biases. Nonetheless, determining the effectiveness of telephone surveys in accurately predicting diabetes risk is valuable for the prospect of assessing large populations as they are cost-effective and quick.

B. Data Attributes

The dataset included a range of attributes offering a comprehensive perspective on factors that potentially influence diabetes risk. These attributes encompassed demographic details, health conditions, and lifestyle habits, each playing a unique role in the analysis.

Demographic factors: Age was classified into 13 categories, representing the adult population's age range from 18-24 years to 80 years and older. Sex was recorded in a binary format.

Lifestyle factors: Smoking status differentiated between individuals who have smoked over 100 cigarettes in their lifetime and non-smokers. Physical activity, excluding job-related activities, was classified as active or inactive in the past month. Fruit and vegetable consumption was recorded based on daily intake versus less frequent consumption. Alcohol consumption, defined differently for men and women, distinguished between heavy drinkers and others.

Health-related factors: High cholesterol status indicated the presence or absence of high cholesterol. Cholesterol checks were recorded, noting whether individuals had been checked in the last five years. Body Mass Index (BMI), a crucial health measure, was included. The occurrence of coronary heart disease (CHD) or myocardial infarction (MI) was noted. High blood pressure (hypertension) was included as well. Self-assessment of health on a scale from excellent to poor (1 to 5) and the number of days with poor mental or physical health in the past month were used to gauge well-being. Mobility challenges, such as difficulty walking or climbing stairs, and a history of stroke were also considered.

These variables collectively provided an in-depth view of an individual's health, thus helping to identify key factors that may indicate a risk of diabetes.

C. Data Cleaning and Preprocessing

The dataset, acquired from Kaggle, was originally sourced from a larger dataset released by the CDC. Before the acquisition, the dataset had already been cleaned, which included the elimination of all null and missing values and underwent transformations to make it more suitable for machine learning. These transformations involved converting complex measurements into binary (sex, smoking, physical activity, fruit intake, vegetable intake, alcohol consumption, high cholesterol, cholesterol checks, heart disease, stroke and difficulty walking) or ordinal features (age, general health) and renaming feature names for enhanced clarity. Continuous variables included BMI, mental health and physical health. There were no duplicates found in the dataset. The target variable, originally in a categorical format with 'non-diabetic' (0), 'pre-diabetic' (1), and 'diabetic' (2) as distinct categories, was altered to a binary format, merging the pre-diabetics and diabetics. This modification, made before the dataset was obtained for this study, was aimed at simplifying the classification task and is justified as a pre-diabetic diagnosis by a doctor indicates a diabetes risk, aligning with the study's focus on identifying risk factors.

The continuous variables were inspected for outliers to ensure data quality and reliability in extracting correlations, values that deviated more than 3 standard deviations from the mean were considered outliers. For BMI, 801 such data points were found. However, upon investigations, these values were found to be valid, the highest being 98 kgm^{-2}, which is a possible value. These records may provide key insights into the relationship between diabetes and BMI and thus were not removed. For mental health, 4373 outliers were found. All values over 29 days of poor mental health fell under this category, however, it would be unreasonable to remove these values as they are entirely valid and may provide valuable information about the relationship between mental health and diabetes.

While preparing the dataset for analysis, two separate features were found to be significantly imbalanced and showed little variance: one for cholesterol checks in the past five years, and the other for heavy alcohol consumption. As shown in Figure 2, both were highly skewed towards one category. This skewness

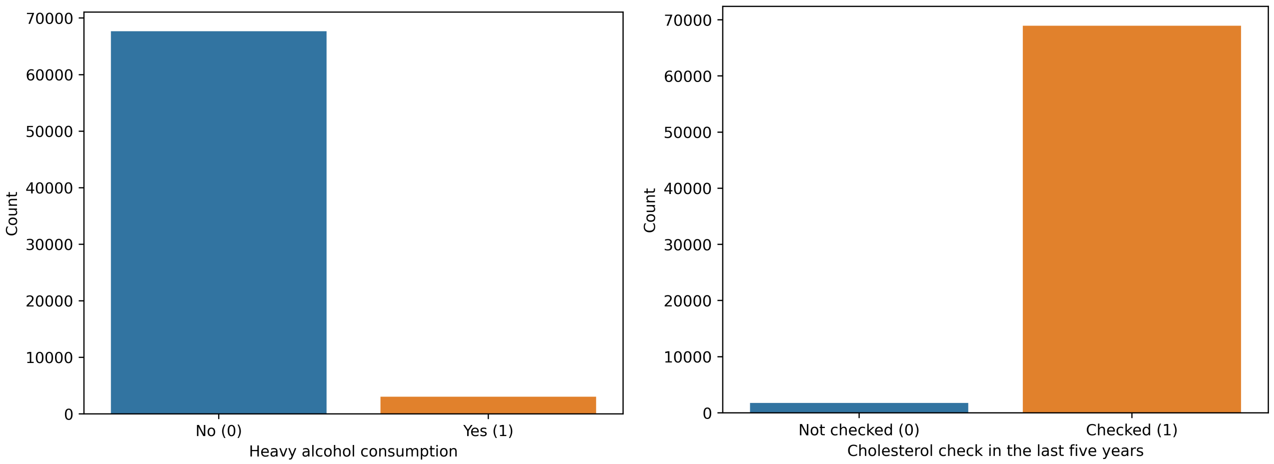

Figure 2: The distribution for the cholesterol checks variable and the heavy alcohol consumption variable were skewed.

could result in poor predictive ability, a likelihood of introducing bias, and a risk of overfitting, which might reduce the model's ability to apply generally. Therefore, these features were not included in the model.

D. Exploring the Features

A correlation matrix was used to understand how feature variables were interrelated (see Figure 3). This was key in identifying potential diabetes predictors. The target variable showed some correlation with high blood pressure (0.38), BMI (0.29), high cholesterol (0.29), and general health (0.41), indicating their importance for classification. Figures 4 and 5 show further exploration into these relationships. General health, physical health, and mental health, being interrelated and conveying similar information, might not all contribute uniquely to the model, and thus may be redundant. Lifestyle factors like smoking, physical activity, and fruit and vegetable intake, while less directly connected to diabetes, were still vital for comprehensive insight. Interestingly, factors such as physical activity and intake of fruits and vegetables are weakly negatively correlated with diabetes, this is crucial when considering prevention in health policy-making.

The analysis was conducted using Python, with specific reliance on libraries such as `Pandas` for data management, `Matplotlib` and `Seaborn` for visualisations, and `Scikit-learn` for ML implementation.

Model	Accuracy	Precision
Naive Bayes	72.5	73.1
Logistic Regression	75.0	74.3
kNN	73.0	72.2
Decision Tree	65.8	67.6
Random Forest	75.0	74.4
SVM	75.4	73.9

Table 1: Performance evaluation metrics for the models implemented to predict diabetes.

III. Results

A. Logistic Regression Analysis

In the selection of a suitable predictive model for this classification task, various factors were considered, including the balance between accuracy, ease of interpretability, and computational efficiency. The process involved testing different models such as naive Bayes, logistic regression, k-nearest neighbour (kNN), decision tree, random forest, and support vector machines (SVM). Table 1 compares performance metrics for these. SVMs performed well but were hard to interpret and took longer to train and use. Ultimately, the decision was made to use a logistic regression model.

Diabetes prediction, a binary classification, fits well with logistic regression, which is tailored for binary outcomes. The model's coefficients, indicating log odds for each unit change in a predictor, are easy to interpret. This clarity is vital in medicine, where understanding each factor's impact is as important as the prediction. This interpretability meets the trans-

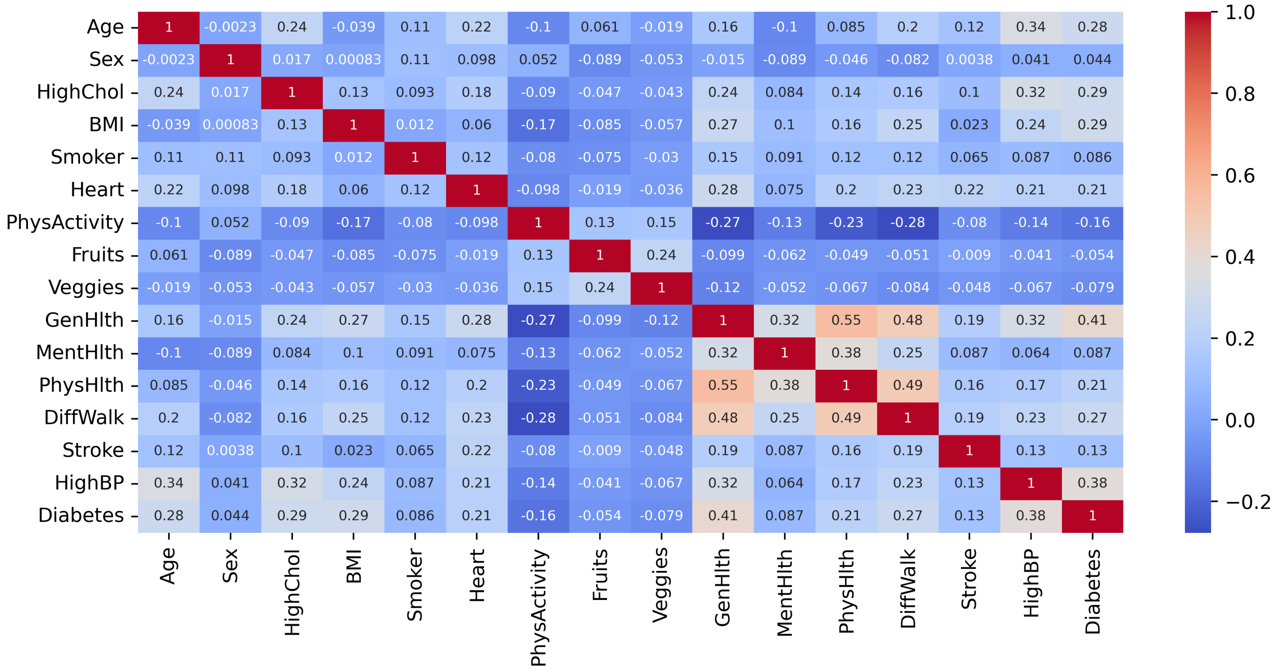

Figure 3: The correlation matrix for the feature variables used. Most variables are uncorrelated. Variables indicating health such as general health, physical, mental health and difficulty walking show some correlation as expected.

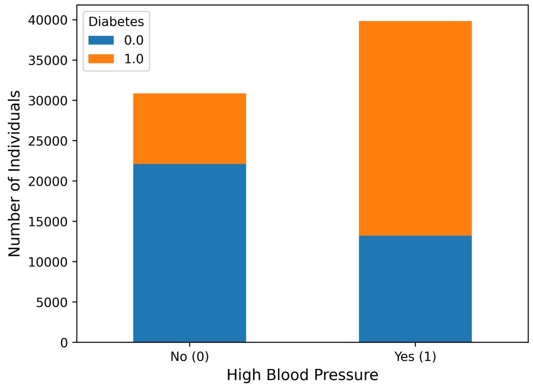

Figure 4: The relationship between high blood pressure and diabetes. A higher ratio of diabetics are found with high blood pressure than not.

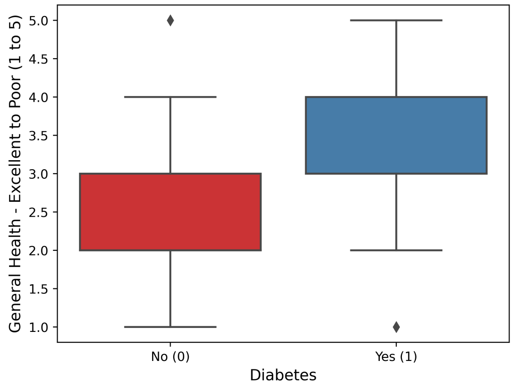

Figure 5: The relationship between general health and diabetes. Poorer general health seems correlated with diabetes risk as expected.

parency needed in healthcare decisions, allowing medical professionals to comprehend and convey the rationale behind predictions. The widespread use and validation of logistic regression in epidemiological and clinical research add credibility and reliability to its use in medical data analysis [13].

Logistic regression, compared to complex models like random forests or SVMs, is computationally simpler, suiting moderate-sized datasets. It models the relationship between a binary outcome (like having or not having diabetes) and independent variables using a logistic function, which keeps the probability output between 0 and 1. In this study, it predicts the likelihood of diabetes or prediabetes based on different risk factors.

The model coefficients reveal how each predictor (like age, BMI, and physical activity) affects the likelihood of diabetes, considering other factors. Positive coefficients mean higher odds of diabetes, and negative ones lower it. The dataset was split 80/20 for training and testing, ensuring sufficient data for inference and reliable model testing.

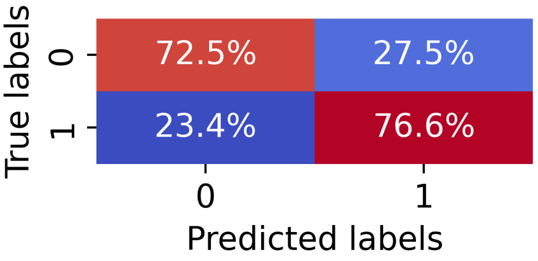

Figure 6: The confusion matrix showing the percentage of accurately classified and misclassfied examples.

B. Model Performance

The model, excluding cholesterol checks and heavy alcohol consumption features, achieved 74.96% accuracy. Accuracy is reliable here, as it measures correctly classified cases against the total tested, and the dataset has a balanced 50-50 target variable. This high accuracy indicates the effectiveness of identified risk factors in distinguishing between diabetics and non-diabetics. Precision, crucial in medical applications, was 0.75, showing the fraction of correctly identified positive cases. While this offers confidence in using these risk factors for diabetes prediction, it may not suffice for the final diagnosis. The confusion matrix, detailed in Figure 6, shows that 77% of diabetic cases were correctly identified.

The confusion matrix shows that 27.5% of diabetics were incorrectly identified as non-diabetic. This misclassification rate makes the model unsuitable for real-world healthcare, as even a small percentage of errors can be significant in large populations or individual diagnoses. It is not recommended for assessing individual health due to the high risk of misdiagnosis. Instead, the model is more appropriate for identifying general trends and estimating statistics at a population level, emphasising the need for human oversight and clinical judgment in healthcare decisions.

The model's predictive ability was assessed with a Receiver Operating Characteristic (ROC) curve, which plots the true positive rate against the false positive rate at various cut-off points of a diagnostic test. The true positive rate indicates the proportion of actual diabetics correctly identified, and the false positive rate shows the proportion of non-diabetics incorrectly classified as having diabetes. These rates are calculated using

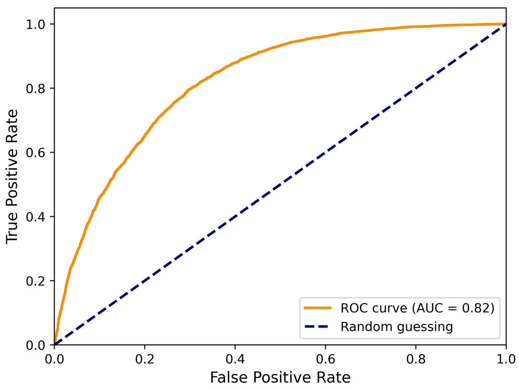

Figure 7: The ROC curve for the logistic regression model for diabetes prediction. The AUC is 0.82, indicating a high predictive accuracy. The dotted line represents a no-skill classifier.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{1}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}, \tag{2}$$

where TP/FP are true/false positives and TN/FN are true/false negatives.

The area under the curve is dubbed AUC, and ranges from 0.5 for random guessing (TPR = FPR), to 1 for perfect differentiation. Here, the AUC was 0.82, indicating a high ability of the logistic regression model to distinguish between patients with and without diabetes.

These performance results prove the model to be fairly reliable for risk factor identification.

C. Important Risk Factors

In logistic regression, coefficients show the importance of each variable for prediction; the higher the value, the greater the influence. To pinpoint the most predictive risk factors for diabetes and suggest a smaller, more efficient set of risk factors for surveying diabetes risk, the most significant features were identified.

To compare the importance of each feature in predicting diabetes, the coefficients were normalised. First, they were converted to absolute values to emphasize their impact size, not the direction (positive or negative). Then, these absolute values were nor-

Feature	Importance
High blood pressure	0.228
High cholesterol	0.192
General health	0.185
Heart disease	0.085

Table 2: The most important features in the logistic regression model in descending order.

malised by dividing them by the sum of all absolute coefficients. This method turns the coefficients into a proportional scale, showing each feature's relative influence among all features. Table 2 details the most important features. A model with just these features yielded a 73% accuracy.

IV. Discussion and Conclusion

Addressing the questions posed in section I, the logistic model indicates that using BRFSS survey data can predict diabetes with about 75% accuracy. This is not sufficient for final diagnoses and medical-grade use, however, it helps assess diabetes risk. Key risk factors identified include high blood pressure, high cholesterol, poor general health, and a history of heart problems, which aligns with current medical research and consensus [14].

High blood pressure and diabetes are closely linked, with people having hypertension being more prone to develop Type 2 diabetes, and vice versa. However, proving direct causation is challenging. Shared risk factors like obesity, poor diet, lack of exercise, and ageing contribute to both conditions, complicating the determination of a direct cause-and-effect relationship [15].

Similarly, high cholesterol is often linked with diabetes due to similar underlying issues like insulin resistance and metabolic syndrome. Metabolic syndrome includes high blood pressure, high blood sugar, excess waist fat, and abnormal cholesterol levels. Insulin resistance, where the body's cells don't respond well to insulin, is crucial in developing Type 2 diabetes and can also cause dyslipidemia, a disorder of lipid metabolism that often manifests as high cholesterol levels [16].

There is a well-established association between diabetes and an increased risk of heart complications such as CHD and MI. Those with diabetes are more prone to CHD and have a higher chance of suffering from MI [17]. Shared risk factors include obesity, poor diet, lack of exercise, and smoking. However, these may not prove to be great predictive factors as diabetes itself can contribute to these heart diseases. The relationship is bidirectional to some extent, with each condition potentially exacerbating the other.

The number of people with diabetes rose from 108 million in 1980 to 422 million in 2014, at a rate that is continually rising [18]. Identifying general health as a diabetes risk factor is key for preventive health policies. Tackling diabetes effectively involves both medical treatment and lifestyle changes. Focusing on overall health, which is closely linked to both physical and mental well-being as shown in Figure 3, can significantly reduce diabetes risks [19, 20]. This understanding can help policymakers globally to encourage lifestyle choices that emphasize cardiovascular health [21].

A simplified model with only four factors accurately predicted diabetes risk at 73%, offering a basis for a concise survey to identify high diabetes risk in populations. This is particularly valuable because it relies on telephone surveys, which are more cost-effective than individual medical tests for every citizen.

One of this study's key limitations was the lack of race data, essential as some ethnic groups have a higher diabetes risk. [22]. Interestingly also, BMI, typically linked to obesity and diabetes risk, was not a major factor, contradicting existing research [23]. Future studies could explore additional variables like insulin resistance, genetics, epigenetics, and erectile dysfunction for more insights. There's also scope for developing a more accurate, medical-grade machine learning model to aid diagnosis, with some models already reaching 90-95% accuracy [24]. Improved models would offer a comprehensive understanding of diabetes risk factors and causes, which will be crucial for preventive strategies in reducing diabetes prevalence.

References

[1] S. Melmed, K. S. Polonsky, P. R. Larsen, and H. Kronenberg, *Williams Textbook of Endocrinology* (Elsevier Health Sciences, 2011).

[2] J. Loscalzo *et al.*, *Harrison's Principles of Internal Medicine*, 21 ed. (McGraw Hill, 2022).

[3] E. D. Parker *et al.*, Diabetes Care **47(1)**, 26 (2023).

[4] S. M. Lyon, I. S. Douglas, and C. R. Cooke, Annals of the American Thoracic Society **11**, 661 (2014).

[5] J. P. Ansah and C.-T. Chiu, Frontiers in Public Health **10** (2023).

[6] M. Roser and H. Ritchie, Burden of disease (https://ourworldindata.org/burden-of-disease), 2021.

[7] CDC, National diabetes statistics report, 2020.

[8] CDC, *BRFSS* (https://www.cdc.gov/brfss/, 2019).

[9] W.H.O, Global report on diabetes (https://www.who.int/publications/i/item/9789241565257), 2016.

[10] U. E. Bauer, P. A. Briss, R. A. Goodman, and B. A. Bowman, The Lancet **384**, 45 (2014).

[11] D. W. Satterfield *et al.*, Diabetes Care **26**, 2643–2652 (2003).

[12] A. Teboul, *Diabetes Health Indicators Dataset* (Kaggle.com, 2022).

[13] E. C. Zabor, C. A. Reddy, R. D. Tendulkar, and S. Patil, International Journal of Radiation Oncology*Biology*Physics **112**, 271 (2022).

[14] Y. Wu, Y. Ding, Y. Tanaka, and W. Zhang, International Journal of Medical Sciences **11**, 1185 (2014).

[15] J. R. Petrie, T. J. Guzik, and R. M. Touyz, The Canadian journal of cardiology **34**, 575 (2018).

[16] J. D. Schofield, Y. Liu, P. Rao-Balakrishna, R. A. Malik, and H. Soran, Diabetes Therapy **7**, 203 (2016).

[17] B. M. Leon and T. M. Maddox, World Journal of Diabetes **6**, 1246 (2015).

[18] W.H.O, Diabetes (https://www.who.int/news-room/fact-sheets/detail/diabetes), 2023.

[19] H. Hamasaki, World Journal of Diabetes **7**, 243 (2016).

[20] C. Garrett and A. Doherty, Clinical Medicine **14**, 669 (2014).

[21] M. Bergman *et al.*, Diabetes management (London, England) **2**, 309 (2012).

[22] J. N. Bottalico, Seminars in Perinatology **31**, 176 (2007).

[23] S. Klein, A. Gastaldelli, H. Yki-Järvinen, and P. E. Scherer, Cell Metabolism **34**, 11–20 (2022).

[24] G. Wagai, S. Firdous, and K. Sharma, Journal of Family Medicine and Primary Care **11**, 6929 (2022).