Journal of High School Science

Machine Learning classifiers for non-invasive questionnaire-based Type 2 Diabetes detection

Lau A.H.C

Submitted: July 10, 2025, Revised: version 1, September 16, 2025
Accepted: September 21, 2025

Abstract

Diabetes is a global health crisis, affecting over half a billion adults worldwide. More alarmingly, nearly half of cases globally, and one in four cases in the US, remain undiagnosed. Untreated diabetes can lead to severe complications such as heart disease, kidney failure, and vision loss, underscoring the need for easily accessible screening methods. Early detection is challenging because diabetes often presents with no symptoms in its initial stages, and many individuals avoid invasive blood tests. This study proposes a non-invasive, questionnaire-based approach to Type 2 Diabetes detection using Machine Learning models. Models were trained on the Diabetes Health Indicators Dataset, derived from a CDC survey, which included questions related to BMI, age, and self-reported health ratings. This method provides a scalable, low-cost screening tool. Three models were evaluated — Logistic Regression, Random Forest, and Neural Network. To minimize false negatives (i.e. missed diabetic cases) due to their severe health consequences, two novel metrics were introduced, *Cost-Weighted Accuracy* and *Cost-Weighted Error Rate*, which incorporated a cost ratio to reflect the higher 'cost' – developing severe health complications - of false negatives, as opposed to the relatively lower cost of false positives. Results showed that the Neural Network model achieved the highest sensitivity (86.66%) and the highest cost-weighted accuracy (75.88%), outperforming the Random Forest and Logistic Regression models.

Keywords

Artificial Intelligence, Diabetes detection, Machine Learning, Diabetes classifier, Logistic Regression, Random Forest, Neural Network, Cost-Weighted Accuracy, Error rate, False negatives

Aydan Ho Ching Lau, Chinese International School, 1 Hau Yuen Path, Braemar Hill, Hong Kong SAR, China.
aydanlau@gmail.com

https://doi.org/10.64336/001c.144808

1. Introduction

Diabetes is a chronic metabolic condition that impairs the body's ability to process blood sugar into energy, leaving affected individuals with dysregulated glucose levels. The condition arises when the pancreas fails to produce sufficient insulin (a hormone regulating blood sugar) or when the body cannot effectively utilize; or resists the effects of; the insulin it produces. Diabetes is broadly classified into two primary types. Type 1 diabetes, an autoimmune condition, results from the destruction of insulin-producing β cells in the pancreas, often diagnosed in children and young adults. Patients of type 1 diabetes require lifelong insulin therapy. Type 2 diabetes, accounting for 90–95% of cases, is linked to insulin resistance and lifestyle factors such as obesity, physical inactivity, and poor diet. It typically develops in adults but is being increasingly observed in younger populations due to rising obesity rates. Unmanaged or undetected diabetes can lead to severe and chronic complications such as cardiovascular diseases (heart attack, stroke), neuropathy (nerve damage causing pain or numbness), nephropathy (kidney failure) and diabetic retinopathy (vision loss). Diabetes also increases susceptibility to infections and presents with non-healing wounds such as foot ulcers, potentially leading to amputations.

The global burden of diabetes has surged in recent decades, underscoring its status as a modern epidemic. According to the International Diabetes Federation (IDF), the number of adults living with diabetes worldwide skyrocketed from 151 million in 2000 to 537 million in 2021—a staggering 256% increase in just two decades (1). This number represents ~10% of the global adult population. This exponential rise is driven by urbanization, sedentary lifestyles, diets high in processed foods, and aging populations.

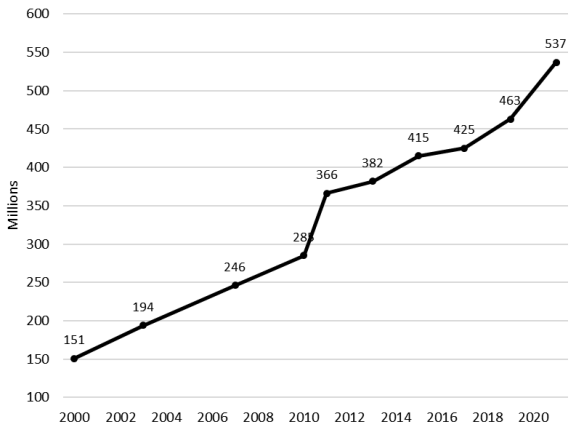

Figure 1. Number of adults with diabetes (millions) over the past 2 decades

This global prevalence underscores the widespread impact of the disease, particularly as it affects a substantial portion of the adult population across urban and rural areas. The IDF projects that this number will rise to 643 million by 2030 and 783 million by 2045. In

the US, the Center of Disease Control and Prevention (CDC) reports that 14.7% of all US adults, or 38.4 million people, were diabetic in 2021 (2).

Table 1. Adults with diabetes and undiagnosed rates by region

Region	Adults with Diabetes (millions)	Predicted by 2045 (millions)	Increase by 2045 (%)	Undiagnosed Rate	Deaths in 2021 (thousands)
North America & Caribbean	51	63	24%	1 in 4	931
South & Central America	32	49	50%	1 in 3	410
Africa	24	55	134%	1 in 2 (54%)	416
Middle East & North Africa	73	136	87%	1 in 3	796
Europe	61	69	13%	1 in 3 (36%)	1,100
South East Asia	90	152	68%	1 in 2	747
Western Pacific	206	260	27%	1 in 2	2,300
Global	537	784	46%	1 in 2 (45%)	6,700

A concerning aspect of diabetes is the high number of undiagnosed cases, which complicates efforts to manage the disease effectively. The IDF reports that in 2021, 240 million adults with diabetes were undiagnosed, accounting for 44.7% of the total 537 million cases (3). Almost one in two adults with diabetes is hence unaware that they have the disease. In the U.S., among 38.1 million diabetic adults, 8.7 million were undiagnosed, resulting in 22.8% of adults with diabetes being unaware of their condition (2). This statistic is particularly alarming, as it means nearly one in four adults with diabetes is not receiving necessary treatment, increasing their risk of developing severe complications.

The importance of early detection cannot be overstated, especially given the potential for non-invasive methods to expand access to screening. Early detection enables prompt intervention and management, which can prevent or delay the onset of serious complications associated with diabetes. Early management, through diet, physical activity, medication, and regular screening, can mitigate these consequences.

In many recent data-driven diabetes detection studies, models are trained using the Pima Indians Diabetes Database (PIDD), a publicly available dataset. This dataset records 768 female instances of Pima Indian heritage in Arizona U.S. with 8 features such as number of times pregnant, blood glucose in an oral glucose tolerance test, triceps skinfold thickness, diabetes pedigree function, BMI, age, blood pressure, and 2-hour serum insulin (4). Despite the extensive use of the PIDD dataset, there are a few concerning issues using the dataset for machine learning application. With only 768 female instances, the small data size, homogeneous ethnicity and female-only data may not accurately represent the actual distribution in a broader population. In addition, the dataset contains significant number of missing/abnormal data (30% of the triceps skinfold thickness data and 49% of the

2-hour serum insulin data are missing) which requires data pre-processing that may skew model predictions (5). The features include costly and time-consuming lab-based results (oral glucose tolerance and 2-hour serum insulin) which may limit the model's application to a wider patient population. Therefore, in this study, a significantly larger dataset with much broader representation was used to address these concerns.

2. Methods

2.1 Dataset

The Diabetes Health Indicators Dataset (DHID) contains healthcare statistics and lifestyle survey information along with their diagnosis of diabetes (6). The DHID was derived from the 2021 Behavioral Risk Factor Surveillance System (BRFSS) by CDC. The CDC, since 1984, has been conducting annual surveys that collect data from adults residing in the U.S. regarding their health-related risk behaviors, chronic health conditions, and their use of preventive services.

The DHID dataset consists of 253,680 instances and 21 features including demographics and answers to life-style questions. The target variable for classification is whether an individual is diabetic/ pre-diabetic, or healthy. The 21 features include mostly yes/no answer to questions regarding high blood pressure, smoking history and frequency, or the consumption of fruit. Out of the 21 features, only 6 questions involve more than yes/no answer; these are: body mass index, rating the participant's health from 1 to 5, age, sex, education level, and income scale. The full list of questions with feature variables are presented in Table 2.

Table 2. Feature variables and questions in the Diabetes Health Indicators Dataset (DHID)

Features	**Questions**
HighDP	Do you have high blood pressure?
HighChol	Do you have high blood cholesterol?
CholCheck	Do you have cholesterol check in last 5 years?
BMI	What is your Body Mass Index?
Smoker	Have you smoked at least 100 cigarettes in your entire life?
Stroke	Have you ever had a stroke?
HeartDiseaseorAttack	Have you had coronary heart disease (CHD) or myocardial infarction (MI)?
PhyActivity	Do you have physical activity in past 30 days?
Fruits	Do you consume fruit 1 or more times per day?
Veggies	Do you consume vegetables 1 or more times per day?
HvyAlcoholConsump	Are you a heavy drinker? (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)
AnyHealthcare	Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc.?
NoDocbcCost	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?
GenHlth	Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 =

	poor
MentHlth	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? scale 1-30 days
PhysHlth	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? scale 1-30 days
DiffWalk	Do you have serious difficulty walking or climbing stairs?
Sex	Are you male or female?
Age	How old are you? 13-level age category 1 = 18-24 9 = 60-64 13 = 80 or older
Education	Education level? scale 1-6 1 = Never attended school or only kindergarten 2 = Grades 1 through 8 (Elementary) 3 = Grades 9 through 11 (Some high school) 4 = Grade 12 or GED (High school graduate) 5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate)
Income	Income scale? scale 1-8 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more

2.2 Feature selection

Feature selection was performed to enhance model performance and interpretability by identifying and removing redundant variables. This process reduces the risk of overfitting, decreases computational cost, and can improve a model's ability to generalize to new data by eliminating noise from highly correlated features.

To initially assess linear relationships between variables, a correlation heatmap was generated (Figure 2). The heatmap did not find any pairwise correlations exceeding a threshold of 0.5. The highest correlation was between the features PhysHlth and GenHlth, which is intuitively reasonable. The absence of strong correlations was likely due to the structure of the dataset. With 16 of the 21 features being binary, the Pearson correlation coefficient is mathematically constrained. Even strongly associative relationships between binary variables often result in modest correlation values, meaning this method alone is insufficient for identifying all forms of redundancy.

To detect more complex multi-collinearity, Variance Inflation Factor (VIF) analysis was performed (Table 3). The initial VIF results identified three features with severe multi-collinearity (VIF > 20): AnyHealthcare, CholCheck, and Education, which were subsequently dropped. A subsequent VIF analysis on the remaining 18 features showed acceptable levels of multicollinearity, with four features yielding moderate VIF values between 5 and 8: Income (7.1), BMI (6.9), Age (6.2), and Veggies (5.8). Given that Income, BMI, and Age are well-established, clinically significant risk factors for type 2 diabetes, they were retained for model training to ensure the retention of critical predictive power, despite their moderately elevated VIF scores. The feature Veggies was dropped and the final set of features with their corresponding VIF values are shown in Table 3.

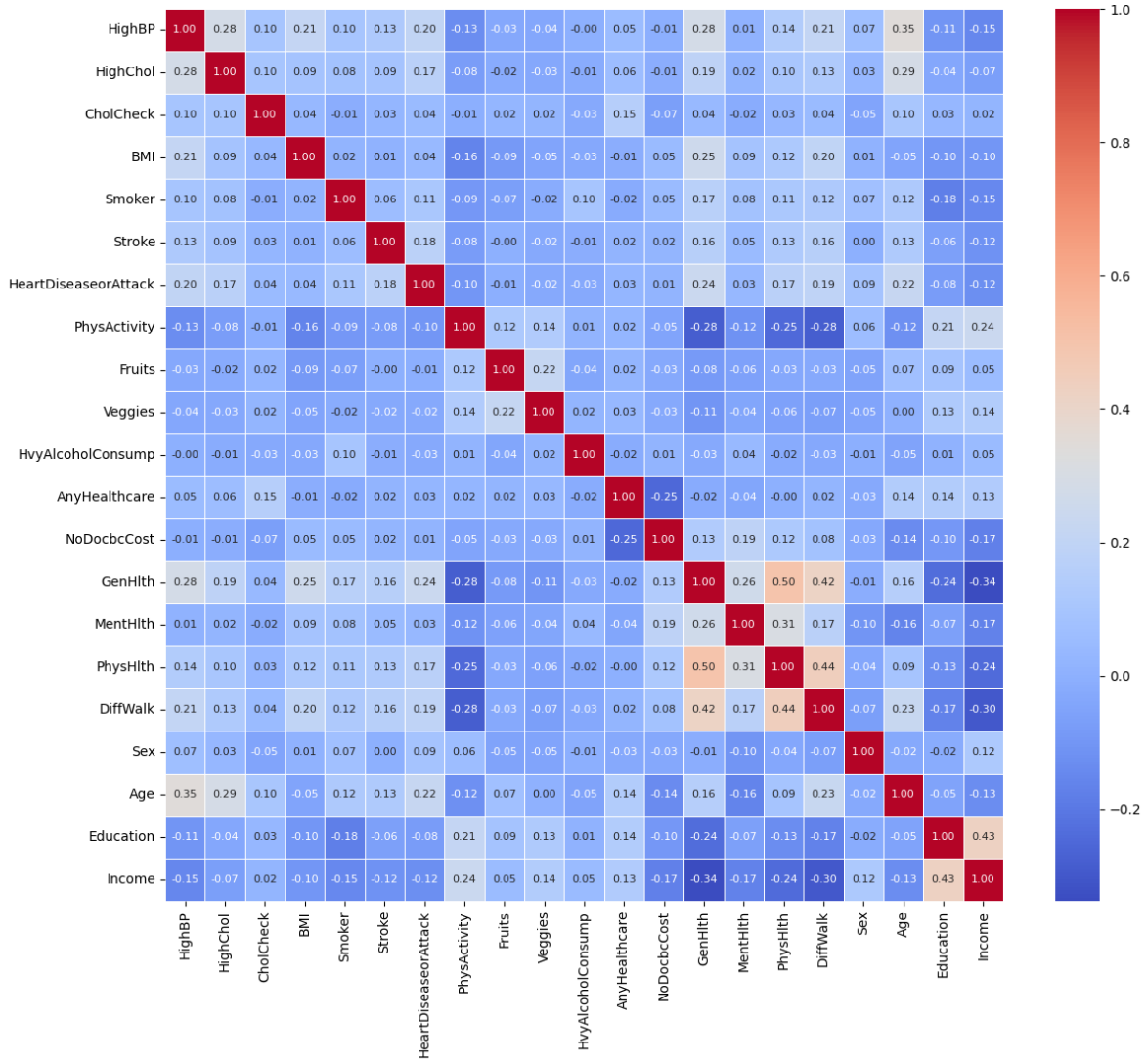

Figure 2. Correlation heatmap of all the 21 features

Table 3. Final VIF scores for the 17 selected features

Feature	VIF Score	Feature	VIF Score	Feature	VIF Score
BMI	6.70	HighBP	2.24	MentHlth	1.49
Income	6.61	Sex	2.01	HeartDiseaseorAttack	1.26
Age	6.11	HighChol	1.94	NoDocbcCost	1.14
GenHlth	4.86	PhysHlth	1.85	Stroke	1.11
PhysActivity	4.59	Smoker	1.81	HvyAlcoholConsump	1.09
Fruits	2.64	DiffWalk	1.74		

2.3 Data pre-processing

The dataset was imported to Python using the Pandas library. The diagnostic result column was encoded as a binary label, with 1 indicating diabetic cases and 0 representing non-diabetic cases. To align with the clinical objective of early detection, pre-diabetic cases were incorporated into the positive (diabetic) class to formulate a binary classification task. This decision was driven by two principal factors. First, the limited number of pre-diabetic instances (~5,600) was deemed insufficient to reliably train a distinct third class. Second, from a clinical perspective, categorizing pre-diabetes with diabetes prioritizes the detection of at-risk individuals, aligning with the objective of early intervention.

All numerical and categorical predictors were normalized to a range of 0 to 1 using the MinMaxScaler in scikit-learn. MinMaxScaler works well to preserve the relationships between data points and is well-suited for features that are naturally bounded. Normalization generally improves the effectiveness of Logistic Regression and Neural Network models.

2.4 Model training and validation framework

To ensure robust performance evaluation, the following framework was adopted:

2.4.1 *Hyperparameter tuning*

For each model, a hyperparameter tuning process with 5-fold cross-validation was performed to ensure optimal performance for each model architecture. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was prioritized as the primary metric for optimization, as it provides a robust, threshold-independent measure of a model's ability to discriminate between classes. If two sets of hyperparameters achieved very similar AUC-ROC, the one with the higher sensitivity was selected since it aligned with the objective of minimizing false negatives.

2.4.2 *Threshold tuning*

For each model with its optimal hyperparameters, the Receiver Operating Characteristic Curve (ROC curve) from the five folds were averaged to generate a single, mean ROC curve. The Cost-Weighted Accuracy (CWA) (defined in section 4) was then calculated across this mean curve for a cost ratio of 10 (discussed in section 5). The probability threshold that maximized the CWA was selected as the optimal operating point for that model, directly aligning the classification decision with the priority of minimizing false negatives.

2.4.3 *Model comparison*

With optimal hyperparameters and thresholds thus defined, the final evaluation was performed. Each model was retrained using its optimal configuration on all five training folds and evaluated on the corresponding validation folds. The results from all five validation folds were averaged, and the final performance metrics such as accuracy, sensitivity, specificity and CWA were calculated and compared.

For the stratified 5-fold cross-validation used above, the diabetic cases were first divided into

5 folds. For each fold iteration, the following procedure was applied, 1] The held-out fold served as the base for the validation set. 2] The remaining four folds of diabetic cases were combined to form the base for the training set. 3] For the training set, non-diabetic cases were randomly selected without replacement and added to the diabetic base to create a 50:50 class ratio, ensuring the model was equally exposed to both classes during training. This approach enhanced the model's ability to distinguish between diabetic and non-diabetic cases, which is critical given the study's emphasis on minimizing false negatives (misclassifying diabetic individuals as non-diabetic) due to the severe health implications of such errors. 4] For the validation set, non-diabetic cases were randomly selected without replacement and added to the held-out diabetic cases such that diabetic cases accounted for 14.7% of the set, preserving the real-world prevalence as reported by the CDC (2) for a realistic assessment of model's generalization ability.

3. Model

3.1 Model introduction

Three types of models were evaluated: Logistic Regression, Random Forest and Neural Network.

Logistic Regression (LR) is a linear model widely used for binary classification due to its simplicity, interpretability, and computational efficiency. It models the relationship between the input features and the target variable using a logistic function, making it an ideal baseline for comparison with more complex models. However, its linearity limits its ability to capture complex, non-linear relationships in the data.

Random Forest (RF) is an ensemble learning method that constructs multiple decision trees during training and aggregates their predictions, typically through majority voting for classification tasks. By combining the outputs of multiple trees, Random Forest reduces the risk of overfitting and is robust to noise in the data. It is particularly effective at capturing non-linear relationships and interactions between features, making it a strong candidate for complex datasets (7).

Neural Network (NN) is highly flexible model capable of learning intricate patterns from data through layers of interconnected nodes. It consists of an input layer, multiple hidden layers that progressively transform the input features, and an output layer that produces the final prediction. This architecture enables the model to capture complex, non-linear relationships in the data, though it typically requires more computational resources and larger datasets to train effectively compared to simpler models.

3.2 Hyperparameter tuning

To ensure optimal performance for each model architecture, a hyperparameter tuning process using 5-fold cross-validation was performed. For the Neural Network, a search was performed over optimizers (Adam, SGD), architectures (5-layer: 64-32-16, 6-layer: 64-32-16-8), dropout rates, and learning rates. The Logistic Regression model was tuned by testing different solvers (such as 'lbfgs',

'liblinear', 'sag', among others) and regularization strengths. The Random Forest was optimized by evaluating the number of estimators, maximum tree depth, and splitting criterion ('gini', 'entropy').

3.2.1 *Logistic Regression hyperparameters*

The default solver 'lbfgs' was chosen as it was suitable for medium-sized datasets. L2 regularization was used with regularization strength set to 1 (C=1). No class weight was applied since the training dataset was already balanced.

3.2.2 *Random Forest hyperparameters*

The selected parameters were 50 estimators, maximum depth of 10 and entropy as the splitting criterion.

3.2.3 *Neural Network hyperparameters*

A sequential model was implemented with five layers: an input layer, three hidden layers (64, 32 and 16 nodes respectively) and an output layer (Figure 3). All hidden layers used the 'relu' activation function. The output layer used a sigmoid activation function to produce binary classification probabilities. The model was trained using the Adam optimizer with a batch size of 32, learning rate of 0.01 and dropout of 0.1.

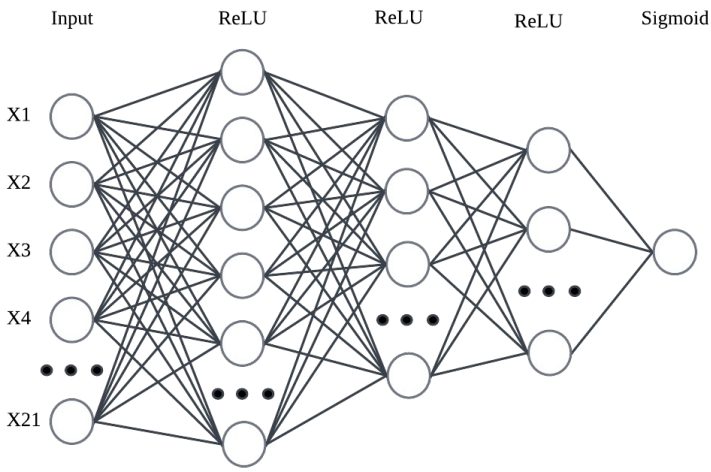

Figure 3. Neural Network illustration

4. Measurement metrics

4.1 Confusion Matrix

Performance metrics are derived from the confusion matrix, which categorizes predictions against actual outcomes, as defined in Table 4. In this context, False Negatives (FN) are defined as those predictions which falsely categorize a diabetic person as being non-diabetic.

Table 4. Confusion Matrix

	Actual Positive (P)	**Actual Negative (N)**
Predicted Positive	True positive (TP)	False positive (FP)
Predicted Negative	False negative (FN)	True negative (TN)

4.2 Performance Metrics

The predictive power of the models was evaluated using accuracy, sensitivity (true positive rate), and specificity (true negative rate), defined as, $Accuracy = \frac{TP+TN}{P+N}$ (eq1), $Sensitivity = \frac{TP}{P}$ (eq2) and $Specificity = \frac{TN}{N}$ (eq3).

4.3 The asymmetric cost of errors – not all errors are created equal

A primary objective was to detect diabetic cases, with a strong emphasis on minimizing false negatives - cases where a diabetic individual is misclassified as non-diabetic. A missed diagnosis can lead to delayed treatment and severe health consequences, a consequence considered far more 'costly' than a false positive, which typically only results in the minor inconvenience and expense of a confirmatory blood test. This priority aligns with the significant medical and economic implications of undetected diabetes.

Given the priority, accuracy (Eq 1) becomes an inadequate measure, as it implicitly assigns equal value to a true positive and a true negative. Similarly, error rate (Eq 4) is also inadequate, as it implicitly treats a false positive and a false negative as equally costly. $Error\ Rate = 1 - Accuracy = \frac{FP+FN}{P+N}$ (eq4).

4.4 New metric – Cost-Weighted Error rate

To address this limitation, a cost-sensitive approach was adopted for threshold tuning and model comparison. This approach prioritizes high sensitivity to minimize false negatives (i.e. maximizing true positives) while maintaining a reasonably low number of false positives by quantifying the trade-off.

This study introduces a cost ratio factored into the error rate to reflect the relative cost implications of false negatives and false positives. The cost ratio (C) is defined in Equation 5 and a new metric, *Cost-Weighted Error Rate (CWE)*, is defined in Equation 6. $Cost\ Ratio(C) = \frac{Cost\ of\ FN}{Cost\ of\ FP}$ (eq5) and $Cost\text{-}Weighted\ Error\ Rate(CWE) = \frac{FP+FN\ x\ C}{N+P\ x\ C}$ (eq6).

The cost ratio is a positive, non-zero number quantifying the relative importance of a false negative versus a false positive. If the errors are equally important, C=1 and Equation 6 reduces to the standard error rate (Eq 4). If a false negative is more important, C>1; if less important, C<1. The "cost" can encompass economic, psychological, physiological, societal and other relevant costs depending on specific use cases. A lower CWE indicates superior performance.

CWE ranges from 0 to 1. A CWE of 0 signifies no errors (no false positives or false negatives), while a CWE of 1 indicates that all predictions

are incorrect (no true positives or true negatives).

4.5 New metric – Cost-Weighted Accuracy

Similarly, the *Cost-Weighted Accuracy (CWA)* can be defined in equation 7 as, $Cost\text{-}Weighted\ Accuracy(CWA)=\frac{TP\,x\,C+TN}{N+P\,x\,C}$ (eq7). It therefore follows that $CWA=1-CWE$ (eq8). A model with a higher CWA demonstrates superior performance for a given cost ratio, analogous to standard accuracy. CWE and CWA measure two sides of the same coin and the choice of which to use depends on whether it is more intuitive to discuss accuracy or error rate in a given context.

In Equation 7, the cost ratio can also be interpreted as a benefit ratio, measuring the relative benefit of a true positive versus a true negative. Generally, the benefit of a true positive can be construed as the cost of a false negative, whereas benefit of a true negative as the cost of a false positive. Thus, Equation 5 can further be expanded as, $Cost\ Ratio(C)=\frac{Cost\ of\ FN}{Cost\ of\ FP}=\frac{Benefit\ of\ TP}{Benefit\ of\ TN}$ (eq9).

Through algebraic manipulation, both the CWA (Eq 10) and CWE (Eq 11) can be expressed in terms of sensitivity (R), specificity (S), cost ratio (C), and class ratio ($p = P/N$). The class ratio is defined as the ratio of actual positives to actual negatives. $Cost\text{-}Weighted\ Accuracy(CWA)=\frac{pCR+S}{pC+1}$ (eq10) and $Cost\text{-}Weighted\ Error\ Rate(CWE)=\frac{pC(1-R)+1-S}{pC+1}$ (eq11).

4.6 Comparison to existing metrics

After optimal hyperparameters were obtained, an average ROC curve was calculated from the five folds. There are several common methods to pick the optimal point on the ROC curve if the objective is to prioritize sensitivity. The easiest method is to target one single sensitivity. For example, if the objective is to correctly identify 90% of all true diabetic cases, then a sensitivity of 90% is chosen. Another method is to find the point on the ROC curve that maximizes the Youden's Index, which combines sensitivity and specificity into one value ($J = Sensitivity + Specificity - 1$) (8). Geometrically, this identifies the point on the ROC curve that is farthest vertically above the diagonal random chance line. A more sophisticated method is to use a Precision-Recall (PR) Curve, which is a plot of precision against sensitivity for all possible thresholds, and select the threshold that yields the highest F1-score (the point closest to the top-right corner on the PR curve) (9).

The primary limitation of these methods is that the trade-off between sensitivity and specificity remains implicit and unquantified in terms of real-world consequences. This trade-off is inherent in every model. A higher sensitivity will always result in lower specificity unless the model is a perfect classifier. It is unclear, for instance, whether a model with sensitivity of 80% and specificity of 60% is preferable to a model with sensitivity of 75% and specificity of 63%, even when prioritizing sensitivity. The implication of gaining 5% in sensitivity at the cost of losing 3% in specificity is difficult to interpret in a clinical setting. Therefore, rather than blithely following the rules of maximizing the Youden's Index or the F1-score,

practitioners need a method to explicitly select a specific trade-off based on the particular application.

The CWA/CWE proposed in this study quantifies this trade-off in an intuitive way. Instead of manipulating abstract percentages, practitioners can decide that they are willing to accept, for example, 5 more false positives to avoid 1 false negative. This decision directly translates to a cost ratio of 5. This clarity allows for the adjustment of the cost ratio to suit specific use cases, moving beyond the "one-size-fits-all" assumption inherent in metrics like the F1-score, the Youden's Index or standard accuracy. For example, a clinic may prefer a cost ratio of 5 while another clinic may choose a cost ratio of 10 as being more appropriate. In addition to tuning the probability threshold of a given model, the CWA can also be used to compare performance across different models for any specific cost-sensitive application.

This study proposed to use both CWA and CWE with a cost ratio of 10, reflecting the judgment that the long-term health and societal costs of one undetected diabetic case is approximately ten times greater than the minor expense and inconvenience of an additional 10 unnecessary blood tests. This number is obviously somewhat subjective. One might argue for any cost ratio from 1 to 15. The discussion section will address how to identify the best model if a different cost ratio is preferred.

5. Results

5.1 Threshold tuning

For each model with its optimal hyperparameters, the ROC curves from the five folds were averaged to generate a single mean ROC curve. The CWA was then calculated across this mean curve for a cost ratio of 10, evaluating probability thresholds from 0 to 1 in increments of 0.01. The threshold that yielded the highest CWA was selected for each model. The plots of CWA against the threshold for the three models are presented in Figures 4, 5, and 6. The optimal thresholds for the Logistic Regression, Random Forest and Neural Network models were determined to be 0.39, 0.42 and 0.44 respectively.

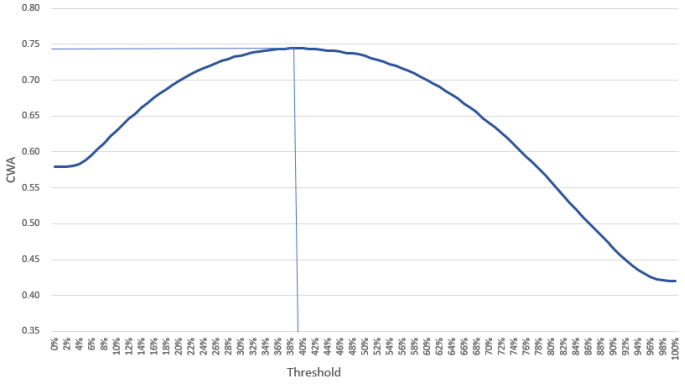

Figure 4. Logistic Regression optimal threshold

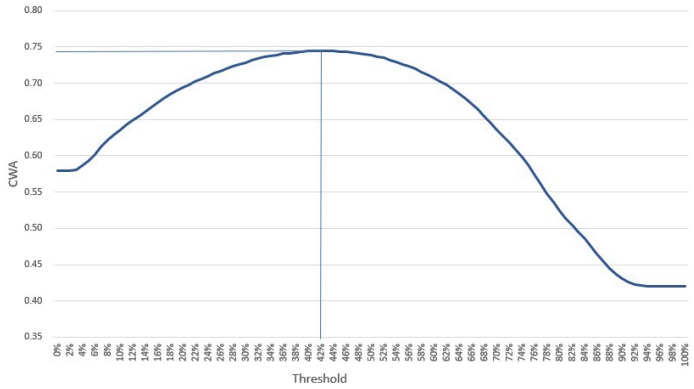

Figure 5. Random Forest optimal threshold

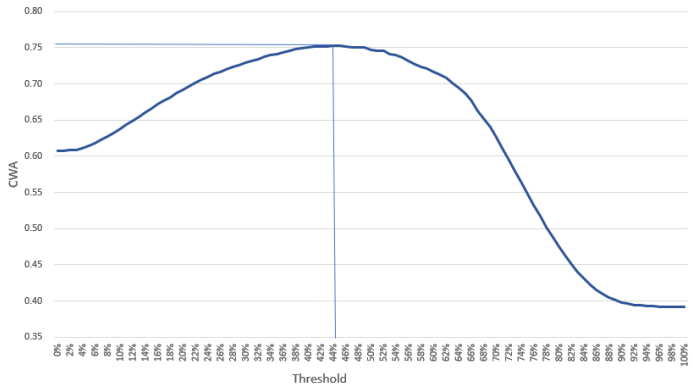

Figure 6. Neural Network optimal threshold

5.2 Model comparison

Table 5. Performance comparison of the three Machine Learning models

Metric	**Logistic Regression**	**Random Forest**	**Neural Network**
Sensitivity	84.85% ± 0.42%	85.03% ± 0.28%	**86.66% ± 1.71%**
Specificity	**60.01% ± 0.24%**	59.99% ± 0.25%	57.31% ± 3.13%
Accuracy	63.66% ± 0.21%	**63.67% ± 0.25%**	61.63% ± 2.42%
CWA (cost ratio=10)	75.73% ± 0.27%	75.83% ± 0.26%	**75.88% ± 0.18%**
CWE (cost ratio=10)	24.27% ± 0.27%	24.17% ± 0.26%	**24.12% ± 0.18%**

With optimal hyperparameters and thresholds thus identified, each model was retrained on all five training folds and evaluated on the corresponding validation folds using its optimal configuration. The results of the three models are shown below. The evaluation metrics included sensitivity, specificity, accuracy, CWA and CWE. These results are summarized in Table 5, with the best results for each metric highlighted in bold. Visualizations are provided in Figure 7 (performance comparison) and Figure 8 (CWA).

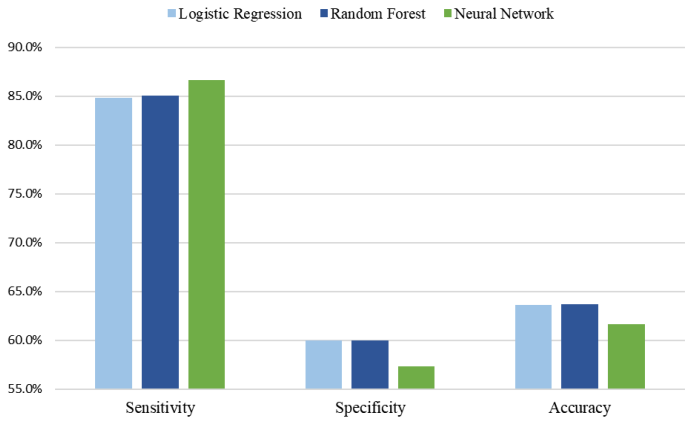

Figure 7. Performance comparison of the three Machine Learning models

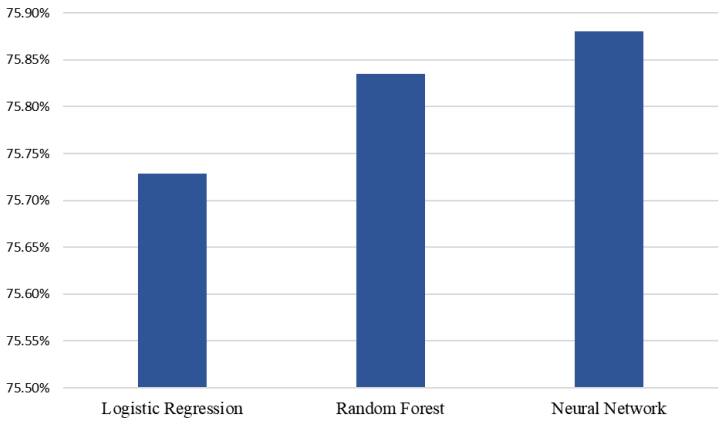

Figure 8. Cost-Weighted Accuracy (CWA) of the three Machine Learning models

6. Discussion

The Neural Network (NN) model achieved the highest sensitivity, correctly identifying 86.66% (± 1.71%) of diabetic cases, followed by the Random Forest (RF) (85.03% ± 0.28%) and the Logistic Regression (LR) models (84.85% ± 0.42%). This suggested that the Neural Network was the most effective model at minimizing false negatives, which was the primary objective of this screening tool. However, this came at a cost to specificity. Logistic Regression led with a specificity of 60.01% (± 0.24%), slightly outperforming Random Forest (59.99% ± 0.25%) and more noticeably outperforming the Neural Network (57.31% ± 3.13%). This demonstrated the fundamental trade-off between sensitivity and specificity inherent in binary classification. As shown in Figure 7, a model that is more aggressive in identifying positive cases (higher sensitivity) will inevitably generate more false alarms (lower specificity).

This trade-off was also reflected in the standard accuracy metrics. The validation dataset was intentionally balanced to reflect the real-world US prevalence of 14.7% diabetic instances. This imbalance meant that true negatives significantly outnumber true positives. Consequently, accuracy, which weights all correct answers equally, tends to follow the trend of specificity. Therefore, while Logistic Regression and Random Forest exhibited the highest accuracy (63.67% ± 0.25%), this metric was not the most desirable measure of performance in this medical context, as it failed to account for the severe consequences of a false negative compared to a false positive.

As a result, the cost-sensitive approach was adopted for model comparison. Using a cost ratio (C) of 10 to reflect the higher personal and societal cost of a false negative versus a false positive, the Neural Network achieved the highest CWA among the three models (75.88% ± 0.18%). Random Forest (75.83% ± 0.26%) ranked second and Logistic Regression (75.73% ± 0.27%) ranked last. However, the absolute performance difference was marginal.

Although the above models were tuned for a cost ratio of 10, their performance can just as well be evaluated using Equation 10 with a different cost ratio but the same sensitivity and specificity. While the Neural Network performed the best at the cost ratio of 10, another model may be superior at a different cost ratio. In fact, calculating the CWA for each model across cost ratios from 1 to 20 found that the Random Forest model outperformed the Neural Network and was the top-performing model when the cost ratio was lower than 9.6. The Neural Network remained the best model for any cost ratio above 9.6. This analysis demonstrated that CWA is a versatile metric for comparing models in cost-sensitive applications, allowing for the selection of the optimal model based on a specific application's risk tolerance.

7. Limitations

While the machine learning models evaluated in this study demonstrated promising results for diabetes detection, several limitations must be acknowledged that could affect their practical deployment and broader applicability.

7.1 US centric dataset

The dataset used to train and evaluate these models was derived from US population, which may not reflect the demographic, genetic, or lifestyle diversity of global populations. Diabetes prevalence and risk factors, such as body mass index (BMI) thresholds or dietary patterns, vary significantly across regions and ethnicities. Therefore, without external validation on international datasets, the models' performance may degrade when applied to non-US populations, limiting their immediate global utility.

7.2 Scaled output

As no model has accuracy of a 100%, it may be misleading to some users if model output is binary – diabetic or non-diabetic. Even with sensitivity of 86.66%, the Neural Network still misses 13.34% of undetected cases. A specificity of 57.31% means 42.69% of non-diabetic persons will suffer the inconvenience and economic costs of being (incorrectly) informed that they may be diabetic.

To mitigate this, a practical implementation should move away from a binary classification and instead output a graduated risk score (e.g., on a scale of 1 to 10). The probability output from the models could be scaled to provide users with a more nuanced understanding of their risk level, labeled from 'very low risk' to 'very high risk'. This approach empowers users to make informed decisions in consultation with healthcare professionals. A clear recommendation could be added, suggesting that individuals with a score above a certain threshold (e.g., 5) should be strongly encouraged to seek a confirmatory medical evaluation.

8. Conclusion

In conclusion, this study showed that Machine Learning models can detect Type 2 Diabetes using non-invasive behavioral and demographic survey questions. Since the screening is non-invasive and can be made easily accessible free-of-charge, such a tool has the potential to help uncover undetected diabetic individuals.

Using the dataset from CDC's BRFSS, the performance of three machine learning models was evaluated. The Neural Network achieved the highest sensitivity of 86.66% and an accuracy of 61.63%. Simpler models, such as Random Forest and Logistic Regression, also performed well with sensitivities of 85.03% and 84.85% respectively.

This paper introduced two novel, related metrics – *Cost-Weighted Accuracy (CWA)* and *Cost-Weighted Error Rate (CWE)* along with the concept of a cost ratio. These metrics are useful in evaluating model performance in scenarios where the cost of a false negative significantly differs from the cost of a false positive, such as the one described in this study. Explicitly quantifying this cost-benefit trade-off, whether in economic or clinical terms, provides a principled method for selecting the most appropriate model for a specific application. The analysis determined that for a cost ratio above 9.60, the Neural Network was the best performing model while the Random Forest was superior for any cost ratio below this threshold. However, all three models performed competitively, suggesting that even a simple Logistic Regression model could capture significant predictive patterns within the data.

This study highlights the potential of machine learning-driven, non-invasive diabetes screening to identify undiagnosed cases. With the global diabetes prevalence projected to rise to 783 million by 2045, accessible screening tools built on these models could play a valuable role in reducing the number of undiagnosed individuals and mitigating the severe long-term complications associated with the disease.

9. References

1. *IDF diabetes atlas 2021*. (2021). IDF Diabetes Atlas. https://diabetesatlas.org/atlas/tenth-edition/

2. *National diabetes statistics report*. (2024, May 15). Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/php/data-research/index.html

3. Ogurtsova, K., Guariguata, L., Barengo, N. C., Ruiz, P. L.-D., Sacre, J. W., Karuranga, S., et al. (2022). IDF diabetes atlas: Global estimates of undiagnosed diabetes in adults for 2021. *Diabetes Research and Clinical Practice*, 183, 109118. https://doi.org/10.1016/j.diabres.2021.109118

4. *Pima indians diabetes database*. (1988). Kaggle. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

5. Wee, B. F., Sivakumar, S., Lim, K. H., Wong, W. K., Juwono, F. H. (2023). Diabetes detection based on machine learning and deep learning approaches. *Multimedia Tools and Applications*, 83(8), 24153-24185. https://doi.org/10.1007/s11042-023-16407-5

6. *Diabetes health indicators dataset*. (2021). Kaggle. https://www.kaggle.com/datasets/julnazz/diabetes-health-indicators-dataset

7. Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32. https://doi.org/10.1023/a:1010933404324

8. Youden, W. J. (1950). Index for rating diagnostic tests. Cancer, 3(1), 32-35. https://doi.org/10.1002/1097-0142(1950)3:1%3C32::aid-cncr2820030106%3E3.0.co;2-3

9. Davis, J., Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning - ICML '06. https://doi.org/10.1145/1143844.1143874