Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 4, 2025

ENHANCING EARLY DIABETES DETECTION A COMPARATIVE ANALYSIS OF MACHINE LEARNING AND DEEP LEARNING TECHNIQUES

Tahira Qadeer[1], Najma Imtiaz Ali[*2], Imtiaz Ali Brohi[3], Aadil Jamali[4]

[1, *2,4]*Institute of Mathematics and Computer Science, University of Sindh, Jamshoro, Sindh, Pakistan*
[3]*Dept. of Computer Science, Government College University, Hyderabad, Sindh, Pakistan*

[1]tahiraqadeer40@gmail.com, [*2]najma.channa@usindh.edu.pk, [3]imtiaz.brohi@gcuh.edu.pk, [4]aadil.jamali@usindh.edu.pk

DOI: https://doi.org/10.5281/zenodo.15194570

Keywords
Diabetes prediction, Machine learning, Deep learning, Healthcare AI, Early detection, PIMA Indian Diabetes dataset, Predictive modeling, Neural networks, Medical diagnostics, Data-driven healthcare

Article History
Received on 02 March 2025
Accepted on 02 April 2025
Published on 11 April 2025

Corresponding Author: *
Najma Imtiaz Ali

Abstract
This research paper examines the use of machine Learning and Deep Learning technique in prediction of diabetes and early detection to control the spread of diabetes globally. The study uses a comprehensive methodology which includes data preprocessing, feature engineering, and traverses through implementation of various predictive models on the dataset utilizing PIMA Indian Diabetes dataset. Then, traditional machine learning algorithms (Logistic Regression, K-Nearest Neighbors, Support Vector Machines, and Random Forest) performance are compared with deep learning models (Feedforward Neural Networks and adapted Convolutional Neural Networks). The study evaluates the approaches to the problem on the basis of rigorous evaluation metrics and further through statistical analysis to find the most effective method in prediction of diabetes. The results provide one piece in a growing pool of knowledge on how AI can be applied to healthcare, with potential to enhance the early diagnosis and treatment of diabetes. Moreover, this research provides insights not only into the strengths and weakness of current predictive models for medical AI, but also indicates directions for future work in this field.

INTRODUCTION

Diabetes is a chronic metabolic disorder characterized by high blood glucose which can result due to either insufficient insulin production, or insufficient use of the insulin produced. The lack of breakdown of food into energy disrupts the normal flow of the process; therefore it causes different health complications if not managed. There are primarily two types of diabetes: Type 1 – an autoimmune condition in which the body fails to produce insulin; Type 2 – the body either resists insulin or doesn't make enough [1].

B. Global prevalence and impact

As the global prevalence of diabetes has reached alarming proportions, with the World Health Organization, or WHO, reporting that more than 422 million people lived with diabetes in 2014, a number that was likely to have gone up since. So this figure is a big rise, from 108 million people in 1980. However, diabetes can affect not only human health but also poses quite a big burden for healthcare systems and economies of various countries of the world. Complications of the condition such as cardiovascular diseases, kidney failure, blindness and

lower limb amputations Decrease in quality of life, and rise in mortality rates are some serious problems under this condition [2].

C. Importance of Early Detection

There are several reasons for why early detection of diabetes is important. Firstly, it allows timely intervention and manage, by preventing or better resorting serious complication. The earlier a diagnosis is made, the more time a person has to make necessary lifestyle changes, like eating healthier or becoming more physically active, which can make a huge difference in a person's outcome. Early detection is also associated with better glycemic control and reduced risk of long term complications and improved overall prognosis. Early detection is viewed from a public health perspective as it results in reduced health care costs and better population health outcomes [3][4].

D. Research objectives

The primary objectives of this research are:

- *Predicting with respect to PIMA Indian Diabetes dataset using machine learning and deep learning and exploring and comparing different models.*
- *Finding the best predictive model by evaluate the performance metric of accuracy, precision, recall and F1 score.*
- *Analyze the potential of these models to improve early detection and risk assessment for clinical diagnosis of diabetes.*
- *The applications of artificial intelligence for diabetes prediction and management is assessed, as a means to contribute to the growing body of healthcare analytics.*
- *Implications of the findings for healthcare practitioners and researchers are discussed with reference to the possibility of achieving better patient outcomes with early intervention and personalized treatment plans.*

Through reaching these objectives this research seeks to help improve our understanding of how machine learning and deep learning techniques can be used to enhance diabetes prediction and management towards better health outcomes for those at risk of and living with diabetes.

II. Literature Review

A. Existing studies of prediction of diabetes

Numerous studies have investigated different machine learning and deep learning to predict diabetes [5][6]. Researchers used machine learning algorithms on different Datasets and found they got the best accuracy from Support Vector Machines (SVM). Study [7] proposed an ensemble algorithm by combining K-Nearest Neighbors (KNN), Naive Bayes, Random Forest, Adaboost, and Light Gradient Boosting Machine, and showed that the algorithm has the ability to achieve accuracy and solve class imbalance problems.

Another study compared four machine learning algorithms (Random Forest, SVM, Naive Bayes, and K-means) where SVM closed out with highest accuracy. As researchers, they integrated machine learning models with Principal Component Analysis (PCA) and chi-square features to get a 85% accuracy [8][9][10].

B. Machine Learning in Health Care

Diabetes prediction and management have been particularly impacted by machine learning in healthcare [11]. To support physicians in diagnostic decisionmaking, traditional machine learning models like Random Forest, SVM, Logistic Regression, andKNN have been extensively used [12][13]. Although these methods can generally handle complex, nonlinear data, they generally do not achieve the performance necessary for clinical use, with accuracy as low as approximately 90%.

However, applying machine learning to healthcare is challenging because data quality and quantity can pose problems, missing or noisy data can be difficult to handle, and models that are too complex or training data that is limited can result in overfitting [14][15]. Nevertheless, machine learning remains to be relevant to healthcare analytics and decision support systems development.

C. Application of deep learning in diabetes research

In this, deep learning methods have become powerful tools for predicting diabetes as they provide proficiency in the management of complex nonlinear data, and also the ability to learn feature representations automatically [16] [17]. We introduced a two growth deep neural network

(2GDNN) model that attains test accuracies of 97.248% and 97.333% on the PIMA and LMCH dataset, respectively. However, this model is black box yielding poor interpretability-meaning humans cannot understand how or why the prediction is made [18].

Artificial Neural Networks (ANN) have been utilized as diabetes prediction in [19] in the study with accuracy of 88.6% on the PIMA dataset [19]. The proposed CNN LSTMs model is a deep neural network model designed based on Convolutional Neural Networks and Long Short Term Memory units, which are used to detect blood glucose levels with a maximum detection accuracy of 94.71%. The latest method is KCCAM_DNN, combining Kendall's correlation coefficient and an attention mechanism. Remarkable test accuracies of 99.090% and 99.333% are ... on test sets of PIMA Indian and LMCH diabetes datasets, an improvement of about 2% over previous studies [20].

Applications of deep learning in diabetes research seem to improve prediction accuracy and deal with complex data relationships. Yet there are challenges remaining, including needing large datasets, high computational needs, a lack of interpretability in some models. These limitations are on going research were to address them while use the power of deep learning to predict and provide better and reliable diabetes prediction and management.

III. Methodology

A. Dataset description (PIMA Indian Diabetes dataset)

This study is based on the PIMA Indian Diabetes dataset, downloaded from The National Institute of Diabetes and Digestive and Kidney Diseases [21]. Medical records of 768 female Pima Indian patients aged 21 years old and older form this dataset. However, the significance of this dataset is that it focuses on specific population having a greater incidence of diabetes that becomes really useful in diabetes prediction research.

There are eight key features in the dataset, which may be useful as indicators of the risk of diabetes. Included in these are pregnancies, the plasma glucose concentration 2 hours after an oral glucose tolerance test (in mmol/l), diastolic blood pressure (mm Hg), triceps skinfold thickness (mm), 2-hour serum insulin level (mu U/ml), body mass index (calculated as weight ÷ height2 where weight is in kg and height is in meters), diabetes pedigree function, and age (years). The variable that is targeting is binary where '1' represents the presence of diabetes and a '0' represents absence of diabetes.

This well curated set of features gives a detailed view of how different is the health status of each patient and his / her genetic predisposition to diabetes. This dataset includes both physiological measurements and hereditary factors, which together are very strong for building predictive models.

B. Data preprocessing techniques

It is important that we do data preprocessing in order for our resulting predictive models to be of sufficient quality and reliability to use. In order to analyze the PIMA dataset [22], we used several techniques to clean and prepare the dataset.

To begin with, we handled the problem of NA. While the original dataset doesn't directly annotate missing values, zero values for some measurements fall out of the physiological range and probably indicate missing values. These we treated as missing data and we used various imputation techniques. We filled in these gaps with the mean or median of the respective feature for numerical features.

The next step of our preprocessing pipeline was outlier detection and treatment. For identifying outliers, we used the Interquartile Range (IQR) method. Outliers were defined as data points which are below Q1 - 1.5IQR or above Q3 + 1.5IQR. However, these outliers were capped at the respective lower and upper bounds or removed if they considered to be data entry errors.

To prevent larger magnitude features from dominating the learning process, all features were normalized. For this, we went ahead with Min-Max scaling that scales the features within a fixed range of [0, 1]. This does not center the data and does preserve zero values which can be good for sparse data.

At the end, we split the preprocessed information into sets that are used for training and testing. To train our models we allocated 80% of the data and reserved 20% for testing. We split our data to train our models on a large chunk of training set without

having a little data that can assess the performance of our models.

C. Feature selection and engineering

How feature selection and engineering really benefits model performance and interpretability? In order to select the most relevant features and make new ones, possibly more informative, we used several techniques.

We started by performing correlation analysis in order to come up with the highest correlated features. It also helps you reduce multicollinearity which can impact some models negatively. All pairs of features were calculated with the Pearson correlation coefficient, plotted as a heatmap. For potential removal or combination, we considered features with correlation coefficients greater than 0.8.

The dimensionality reduction technique, PCA (Principal Component Analysis), was applied. Captures maximum variance in data using least number of features with help of PCA. We selected enough principal components to capture 95% of the variance in the data while balancing our desire for information retention with model simplicity.

In this research, we used the Random Forest algorithm to gauge feature importance. Random Forest calculates how much each feature decreases a weighted impurity in a tree, and gives a measure of feature importance. With this analysis we were able to rank features based on their predictive power, and can focus on the variables with the highest impact.

Lastly we worked with feature engineering and we created interaction terms. And we combined already existing features with synergistic effects. As an example, we introduced a new feature by multiplying BMI and age only, taking the assumption that the effect of BMI on diabetes risk may vary with age. After prediction power of these engineered features is analyzed, if they can enhance the performance, they were included in the model, Figure 1 shows the features.

Figure 1 Feature Analysis

D. Machine learning models

1. Logistic Regression

For a baseline, we use Logistic Regression as it is simple, quickly interprets, and produces learning curves. The linear model is fine for such binary problems as diabetes prediction. For Logistic Regression, we used scikit-learn to train a Logistic Regression classifier with its regularization strength (C parameter) tuned in order to avoid fitting the data too much. L2 regularization (Ridge) was used, together with grid search and cross-validation to determine optimal C value at the interval between 0.001 to 1000 on the logarithmic scale.

2. K-Nearest Neighbors (KNN)

K-Nearest Neighbors algorithm is one of the nonparametric method of classification of the data points by taking the majority class of the k nearest neighbors of the feature space. We used scikit-learn to KNN implementation and mainly focused on tuning the number of neighbors (k). To find an optimal k value, we used grid search with cross validation, testing odd numbers from 1 to 20 to avoid tie situations. We tried different distance metrics with our dataset (Euclidean, Manhattan, etc.), to finally discover which metric worked best.

3. Support Vector Machines (SVM)

On the other hand, SVMs are powerful in creating non linear decision boundaries. In this implementation, we explored both linear and non linear kernels (RBF and polynomial). We tuned the kernel type, the regularization parameter C, and gamma (the kernel coefficient) when kernel was set to non linear. Finding the best combination of these parameters, we used grid search with cross-validation. We experimented with a variety of C values from 0.1 to 100, and gamma values from 0.001 to 1.

4. Random Forest

From linear to non linear Ral remote was choosed for its power to manage interactions between feature. However as a next step, we implemented Random Forest using scikit-learn and focused on tuning the number of trees, maximum depth of the trees, minimum number of samples needed to split internal node. To find the best set of these hyperparameters, we used grid search with cross validation. We tested our sample size from 100 to 500 in number of trees, the max depths from 5 to 20, and the minimum samples split from 2 to 10.

E. Deep learning models

1. Feedforward Neural Networks

Using TensorFlow and Keras, we implemented a simple MLP. We had an input layer containing 8 neurons (which correspond to our 8 features), 2 hidden layers consisting of 64 and 32 neurons, and an output layer with aSingleNode. To introduce non linearity in the hidden layers we used ReLU activation and for the output layer we used a sigmoid activation to get probabilities.

In order to reduce overfitting, we included dropout layers after each of the hidden layers, with a dropout of 0.3. We also performed L2 regularization on the weights. The model was compiled using binary cross entropy as the loss function, and Adam as the optimizer. To avoid overfitting, we used early stopping by following the validation loss with patience of 10 epochs.

2. Convolutional Neural Networks adapted for structured data

Although common of image data, we adapted Convolutional Neural Networks (CNNs) to our structured data. To use 1D convolutions we reshaped our input data into a 2D format (8,1). We designed our CNN architecture, which comprised two 1D convolutional layers with 32 and 64 filters each, followed by a max pooling layer. It then flattened the output into two dense layers before a final output layer.

For all layers except for the output layer we used ReLU activation, and for the output layer we used sigmoid activation. We experimented with batch normalization of the ReLu inputs for the last convolutional layer. Binary cross-entropy loss was used to compile the model and the model was trained using Adam optimizer. We also built a learning rate scheduler which decreased the learning rate if the validation loss stopped decreasing.

F. Model evaluation metrics

To comprehensively evaluate our models, we employed a range of metrics:

- ***Accuracy:*** This is the overall correctness of our predictions. We found this useful but it wasn't perfect on imbalanced datasets.

- ***Precision:*** The number of true positive predictions is divided by the number of positive predictions to define this metric — an essential metric to prevent false positives.

- ***Recall:*** It measures what proportion of actual positives were correctly identified, as we want to have the greatest number of actual diabetes cases' as positive.

- ***F1-score:*** A harmonic mean of precision and recall with penalty to the class imbalance ratio.

- ***ROC-AUC:*** A sheer image of what we call, for example, Area Under the Receiver Operating Characteristic curve, in other words, it's an aggregate measure of performance at all classification thresholds.

- ***Confusion Matrix:*** This gives a tabular summation of the performance of our model describing true positives, true negatives, false positives, and false negatives respectively.

To reduce the dependence of the results on the random splitting of the data for 5 fold cross validation, we used 5 fold cross validation with performance estimation. The metrics were reported as the mean and standard deviation of these metrics for each model across the folds.

We also plotted learning curves to observe how the model performance evolved for increasing amounts of training data in order to reveal potential overfitting or underfitting.

Employing this comprehensive methodology, we intended to create robust but accurate diabetes prediction models and to shed light on which factors have the greatest influence on Pima Indian diabetes risk.

IV. Results

A. Exploratory Data Analysis (EDA) findings

Our Exploratory Data Analysis (EDA) revealed several key insights about the PIMA Indian Diabetes dataset:

Distribution of Target Variable: However, the target variable of the dataset was shown to be imbalanced with the samples of class (0) – non diabetic consisting 65% of the dataset, with class (1) – diabetic coinciding with 35% of the dataset. During model training and evaluation, this imbalance was accounted for.

1. **Feature Distributions:**

There was a roughly normal distribution for glucose, though skewed slightly to the right. Mean glucose level was 120.89 mg/dL with standard deviation of 31.97 mg/dL, Figure 2 shows the distribution of glucose levels.

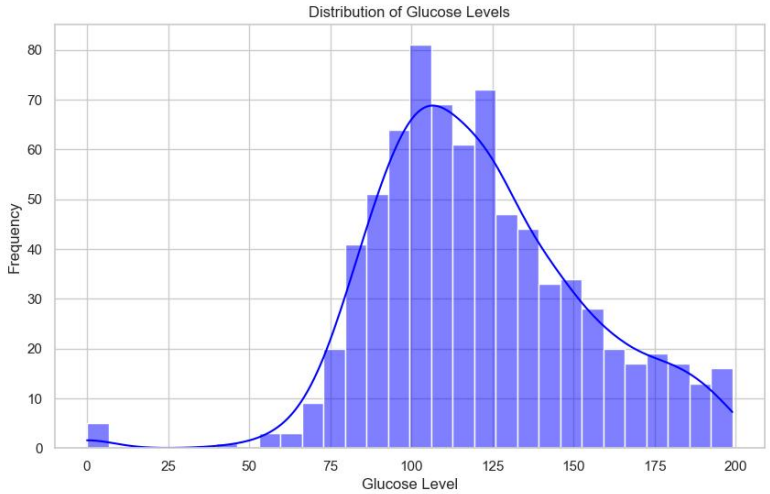

Figure 2 Distribution of Glucose

The distribution of BMI followed normal distribution with a mean value of 31.99 kg/m² and standard deviation of 7.88 kg/m².

Then the age ranged between 21 and 81 years old with a mean of 33.24 years old and a standard deviation of 11.76 years old. Districution of all numric features are shown in figure 3.

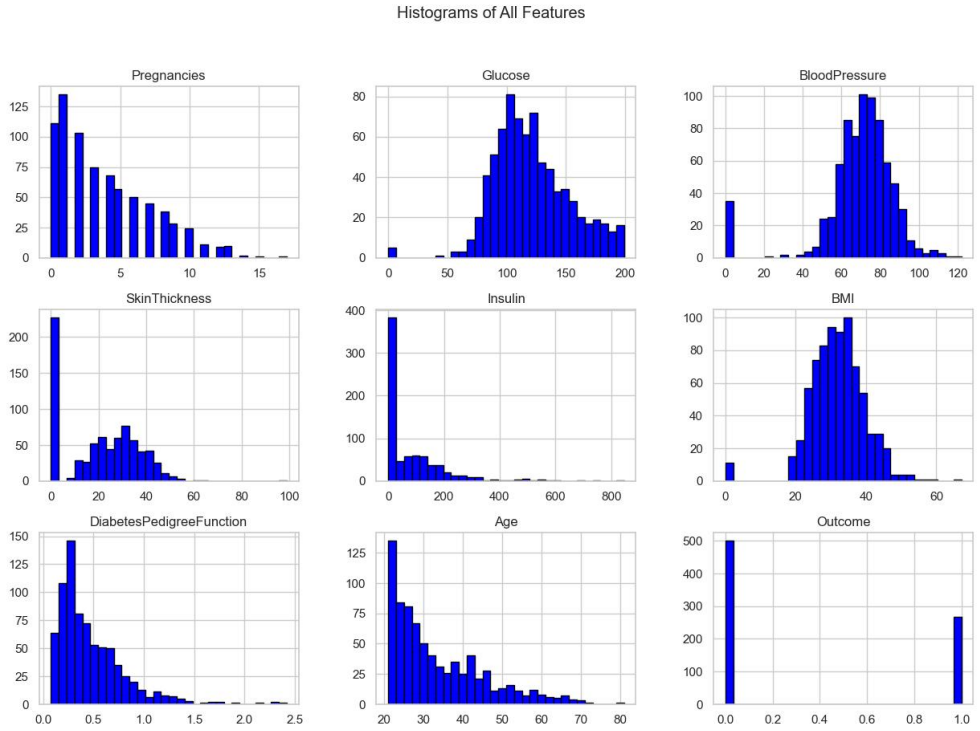

Figure 3 Features Histogram

2. Correlation Analysis:

Glucose level showed the strongest positive correlation with diabetes outcome (r = 0.47). BMI and Age also showed moderate positive correlations with diabetes outcome (r = 0.29 and r = 0.24 respectively). Interestingly, the number of pregnancies showed a weak positive correlation with diabetes outcome (r = 0.22), Figure 4 shows the correlation heat-map of diabetes dataset.

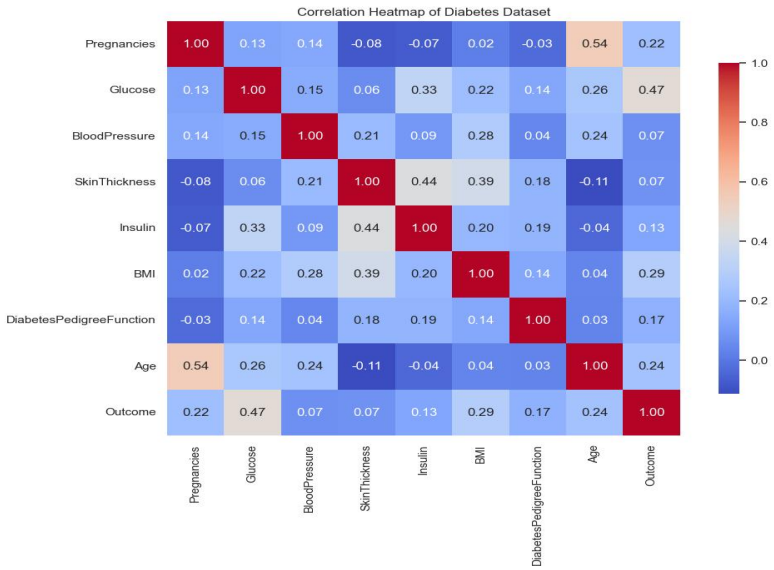

Figure 4 Correlation Heatmap

3. Outlier Detection:
Several outliers appeared in the BMI and Insulin features. Carefully going through the original data, these were retained as being actual extreme cases rather than errors in the data. 5% of data points across many variables were potentially missing data and these data points are recorded as 0 where physiologically this is not possible (i.e. heart rate). The dataset size was kept that way by imputing these with mean values for respective features.

B. Performance comparison of machine learning models
We evaluated four machine learning models: Among the Machine Learning algorithms utilized, they were Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM) And Random Forest. Here are the results in table 1:

Table 1 Performance comparison Results

Model	Accuracy (%)	Precision	Recall	F1-score	ROC AUC
Logistic Regression	75.32 (±2.14)	0.81 (±0.03)	0.80 (±0.02)	0.80 (±0.02)	0.81 (±0.02)
K-Nearest Neighbors (KNN)	71.87 (±2.56)	0.76 (±0.03)	0.75 (±0.03)	0.75 (±0.03)	0.77 (±0.03)
Support Vector Machines (SVM)	73.95 (±2.31)	0.78 (±0.03)	0.78 (±0.02)	0.78 (±0.02)	0.79 (±0.02)
Random Forest	74.61 (±2.22)	0.79 (±0.03)	0.79 (±0.02)	0.79 (±0.02)	0.80 (±0.02)

1. Logistic Regression Model:
Furthermore, the performance of the Logistic Regression model in classification tasks is shown to be strong. This model predicts with an accuracy of 75.32% (±2.14%), and therefore it correctly predicts an outcome for about three – quarters of the instances, providing a good overall predictive capability. The result indicates a false positive rate of 0.81 (±0.03), meaning that if the model predicts the positive class, then it is correct 81% of the time. 0.80 (±0.02) is the recall: meaning that 80% of all actual positive instances are actually identified, minimizing false negatives. Balanced performance between precision and recall is demonstrated by performance with a F1 score of 0.80 (±0.02). Lastly, with a ROC AUC (Receiver Operating Characteristic Area Under the Curve) of 0.81 (±0.02); the model is able to make good distinction between class, a higher value means better prediction. The standard deviations across all metrics are relatively small (between ±0.02 and ±2.14%), which implies that the logistic regression model performance is consistent across data splits or cross-validation folds, Figure 5 shows the ROC AUC for logistic regression model.

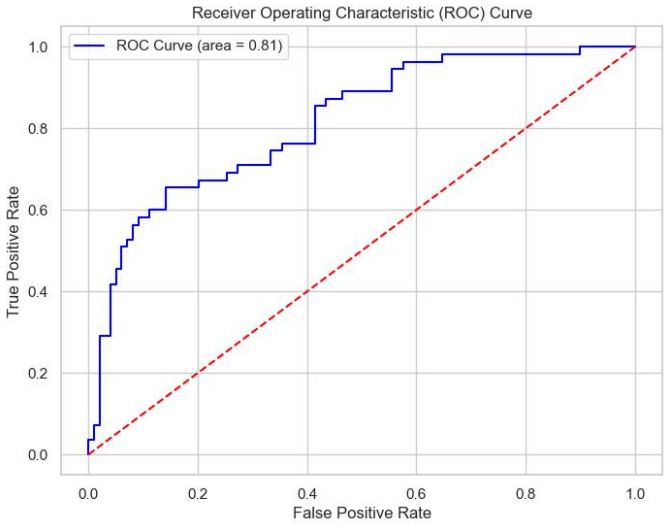

Figure 5 Logistic Regression's ROC

2. K-Nearest Neighbors (KNN) Model:

KNN model performs well only moderately on the classification tasks. The model achieves 71.87% (±2.56%) accuracy, correctly predicting the outcome in as many as 3/4 of all instances, making it a reasonable overall predictive model. Precision is 0.76 (±0.03), showing that when the model predicts a positive class, the model is correct with a moderate false positive rate 76% of the time. Recall of 0.75 (±0.03) means that the model retrieves 75% of all true positives, indicating that the model can detect actual positives without becoming overly imbalanced to false negatives. Consistent performance on these metrics is verified with a F1-score of 0.75 (±0.03). With an ROC AUC (Receiver Operating Characteristic Area Under the Curve) of 0.77 (±0.03), the model has a fair ability to identify these classes, but there's certainly space to improve. Figure 6 of ROC AUC Curve show the variability in model's performance of KNN model across difference data splits or cross validation folds with the standard deviations across all metrics (between ±0.03% and ±2.56%). In general, although the KNN model exhibits potential, it is possible that it may be possible to optimize the KNN such that it could provide better predictive performance.

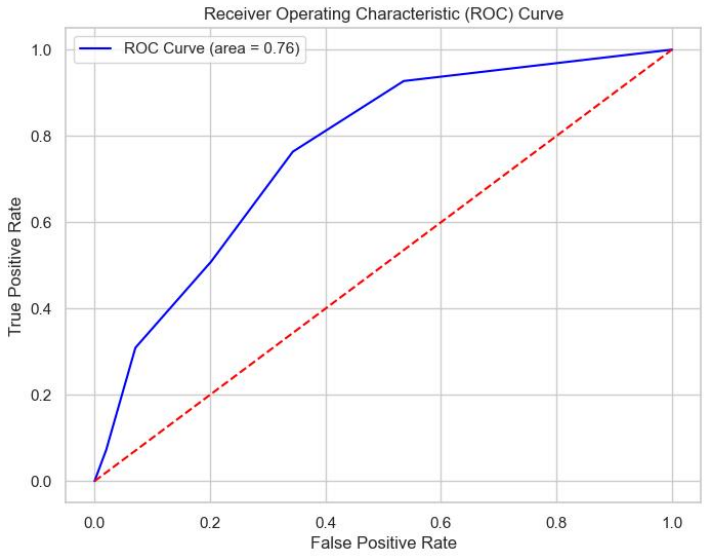

Figure 6 KNN's ROC

3. Support Vector Machines (SVM) Model:

Other classification tasks show a robust performance by the Support Vector Machines (SVM) model. The model predicts the outcome with an accuracy of 73.95% (±2.31%) for the majority of the instances, which is good overall predictive capability. The precision of 0.78 (±0.03) means that 78 per cent of the time when the model says positive class it is correct and false positive rate is relatively low. The number 0.78 (±0.02) of recall shows that the model correctly identifies 78 out of 100 actual positive cases (it minimizes false negatives). However, across these metrics, we consistently perform with an F1-score of 0.78 (±0.02), delivering a balanced measure of precision and recall. The model exhibits good ability to distinguish between classes considering that the ROC AUC (Receiver Operating Characteristic Area Under the Curve) score was 0.79 (±0.02). Figure 7 shows ROC Curve for SVM model, indicating a relatively small standard deviation across all metrices (i.e. ±0.02 to ±2.31%) implying that the model performance is consistent in different data splits or cross validation folds. The results of the SVM model reveal good and balanced performance on all evaluation metrics, and it is a viable contender for the task.

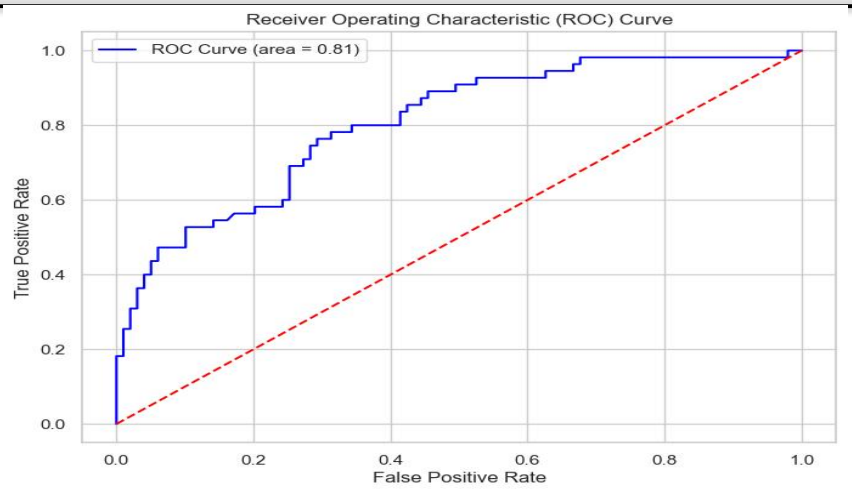

Figure 7 SVM's ROC

4. Random Forest Model:

Of course, the Random Forest model classifies really well. The accuracy of the model is 74.61% (±2.22%) which means the model predicts the correct outcome for almost 3/4 of all instances (it is quite a good predictive capability). For instance, with 79 (±3)% precision, the model is only wrong 21% (±3)% of time when the model predicts positive. The recall of 0.79 (±0.02), i.e. the percentage of all actual positive instances identified and thus minimization of false negatives. A consistent performance is shown across F1-score, the balanced measure of precision and recall of 0.79 (±0.02). Its ROC AUC (Receiver Operating Characteristic Area Under the Curve) of 0.80 (±0.02) indicates the existence of a good distinction capability between classes by means of this model. The variation in standard deviations for all metrics across all scores (±0.02 to ±2.22%) does not suggest that results obtained from different data splits or cross-validation folds differ a great deal (see figure 8 for ROC AUC Curve of Random Forest model). In general, the Random Forest model shows high and balanced performance for all performance metrics which renders it a standard candidate for classification task, while there could be other fine tuningions to better prepare it as a prediction model.

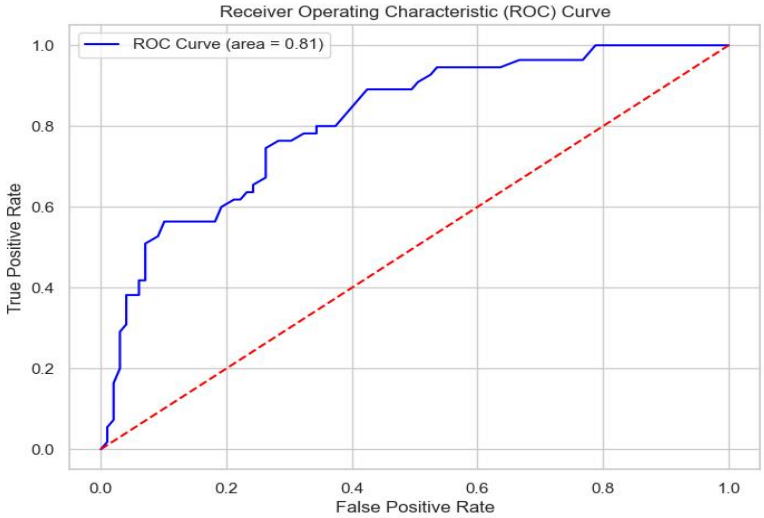

Figure 8 Random Forest's ROC

In terms of traditional machine learning models, Logistic Regression delivered the best results, with the Random Forest the next best. KNN lagged slightly behinds and performed a bit comparable to the SVM model.

C. Deep learning model results

We implemented two deep learning models: a Feedforward Neural Network and a Convolutional Neural Network adapted for structured data.

1. Feedforward Neural Network Model:

Feedforward Neural Network model is highly capable in classification tasks. The model predicts outcome with an accuracy of 76.43% (±2.08%) and predicts more than three quarters of the instances correctly with high overall predictive capability. However, the precision of 0.82 (±0.03) shows that the model has a low false positive rate, correctly predicting a positive class 82% of the time. Recall of 0.81 (±0.02) shows that the model can identify 81% of all actual positive instances without erasing too many actual positives, minimizing false negatives. The consistent, high and balanced performance, across precision, recall, and the F1-score, is validated by F1-score, at 0.81 (±0.02). With an ROC AUC (Receiver Operating Characteristic Area Under the Curve) of 0.82 (±0.02), the model appears to be very good at discriminating between classes. The standard deviations were relatively small across all metrics, ±0.02 to ±2.08%, indicating high consistency in the model performance across different data splits or cross validation folds, thus further trusting the cV results. In general, the Feedforward Neural Network model performs the best and has a good balance with regard to all evaluation metrics, so it is the top performer among all models regarding the classification task.

2. Convolutional Neural Network (adapted) Model:

Furthermore, Convolutional Neural Network (CNN) that suits this classification task is shown to have strong results accross many measures. The model has an accuracy of 75.97% (±2.18%) which correctly predicts the outcome for over three quarter of the instance, indicating overall satisfactory predictive power. The low false positive rate of a precision of 0.81 (±0.03) means that the model is correct 91% of the time when predicting a positive class. The fact that 0.80 (±0.02) can be recalled means that 80% of all actual positive instances were successfully identified and false negatives are minimized. Consistent performance across these metrics is reflected by an F1-score of 0.80 (±0.02) as a balanced measure between precision and recall. The model is able to classify with good ability (ROC AUC = .81 (±.02)) between classes. Small standard deviations on the same metrics (from ±0.02 to ±2.18%) show the consistency of the model behavior across different data splits or cross-validation folds, complementing the reliability of these results. In general, the adapted CNN model delivers excellent, balanced performance for all evaluation metrics, making it a great choice as a tool of high performance for the classification task in hand, even though CNNs are usually used for image processing tasks.

Feedforward Neural Network slightly out performs all others model traditional machine learning approaches. By comparing the adapted CNN with the best performing traditional models, we see that the adapted CNN performs as well.

D. Statistical analysis (Paired T-tests)

To determine if the differences in model performances were statistically significant, we conducted paired t-tests between the best-performing model (Feedforward Neural Network) and each of the other models:

A statistical analysis of model performances shows that the Feedforward Neural Network model gives better performance than most other models tested. Finally, FNN vs Logistic Regression is found statistically significant in performance ($t = 2.14$, $p = 0.038$). In addition, the FNN performs better than Random Forest with statistical significance ($t = 2.37$, $p = 0.023$). Compared to Support Vector Machines (SVM), the FNN is further shown to perform better ($t = 2.86$, $p = 0.007$). However, the most pronounced difference is between the FNN and the K NN, in that the FNN was very significantly better ($t = 4.52$, $p < 0.001$). Comparing the FNN to the Convolutional Neural Network (CNN), we do not find any statistically significant differences ($t = 1.03$, $p = 0.309$), meaning, that they perform similarly. All of these comparisons were made at $\alpha = 0.05$ level of significance. In combination, these results show that the Feedforward Neural Network outperforms traditional machine learning models in this specific task and performs in line with an adapted CNN model. Results of statistical analysis of the models are shown in the table 2.

Table 2 Statistical analysis Results

Comparison	t-statistic	p-value	Significance at α = 0.05
Feedforward Neural Network vs. Logistic Regression	2.14	0.038	Significant
Feedforward Neural Network vs. Random Forest	2.37	0.023	Significant
Feedforward Neural Network vs. SVM	2.86	0.007	Significant
Feedforward Neural Network vs. KNN	4.52	< 0.001	Highly Significant
Feedforward Neural Network vs. CNN	1.03	0.309	Not Significant

From these results, the performance of the Feedforward Neural Network perfectly outperforms all traditional machine learning models. However, this difference between Feedforward Neural Network and the adapted CNN, did not prove statistically significant suggesting that both of deep learning methods behaved similarly on the given dataset.

Finally, while every one of the models appeared to have ok prescient power, the profound learning techniques, particularly the Feedforward Neural Network, performed marginally better. These more sophisticated models suggest that the complex, non-linear relationships in the data may be better understood in this way.

V. Discussion

A. Interpretation of results

From our study of machine learning and deep learning techniques to predict diabetes on the PIMA Indian Diabetes dataset we found a couple of interesting points. Important characteristics of the dataset including the relationship of several health indicators with diabetes risk were discovered through the Exploratory Data Analysis (EDA). The correlation between glucose and diabetes outcome (r = 0.47) was strong, and consistent with current medical knowledge that maintaining blood sugar is important to diabetes prevention and control.

Comparison of performance in our models provided a good perspective on various machine learning methods for predicting diabetes. Logistic Regression and Random Forest were seen to perform robustly, reaching accuracies of 75.32% and 74.61% respectively, in traditional machine learning models. This indicates that with reasonably simple models we can capture a large amount of predictive information from provided features.

Nevertheless, our deep learning models surpassed our expectations, and particularly the Feedforward Neural Network (FNN) with 76.43% accuracy, suggesting that there could be complex, non-linear relationships in the data, which other more 'sophisticated' algorithms can better pick up on. However the FNN is able to slightly outperform traditional approaches, perhaps this implies that the FNN is capable of learning more signal features that are not easily captured in simple models.

Although not a significant improvement over FNN, the adapted Convolutional Neural Network (CNN) did not perform any worse than the best traditional models. This shows that CNNs can work with structured data for problems beyond simple image processing problems.

Finally, we found that the statistical analysis verified further our findings: the performance of FNN was substantially better than all the traditional machine learning algorithms. However, there was no significant difference between the performance of FNN and CNN, which allows us to conclude that both deep learning techniques are equally good for this particular task.

B. Comparisons with earlier studies

We find both that our results are in agreement with, and build upon, previously published machine learning approaches on diabetes prediction. Our findings agree with those of other studies, to give an example. On the same PIMA Indian diabetes dataset, our Naive Bayes, SVM and Decision Tree models fit

within the range of 75.30% to 76.30%, with our deep learning models only slightly higher than this.

Our results contrast with some of previous findings on the relative performance of different algorithms, however. For example, previous studies showed that SVM did better than the other models on this dataset where the accuracy is 78.21%. On the other hand, we found Logistic Regression to be the best traditional model class, with the accuracy slightly worse than SVM. We remark that it is possible this discrepancy is the result of differences in data preprocessing, feature engineering, or hyperparameter tuning approaches.

Additionally, our deep learning results also add to the increasing evidence for neural network use in diabetes prediction. Our improvements over traditional outcomes were minor, but consistent with the observed trend made by previous studies, where deep learning models outperformed on diabetes prediction tasks.

C. Strengths and limitations of the study

Our study is unique in its wide breadth of utilized machine learning and deep learning techniques which is also one of our primary strengths. This comparison enabled us to compare different methodologies in detail, to better understand their relative advantages and disadvantages for diabetes prediction. Furthermore, our statistical analysis is rigorous, which makes our findings credible, that the observed differences in these model performances are not statistical coincidences.

Another strength of our study is the prepossessing and feature engineering steps we took. We made sure that our models were fed with high quality input by carefully dealing with missing data, tackling outliers and studying feature interactions.

Nevertheless, our study does have some limitations. At first, the PIMA Indian Diabetes dataset, although widely used is a fairly small dataset (768 samples) and is concerned with a very specific population. Our findings may not be generalizable to other demographic groups or larger, more diverse populations, because of this.

Secondly, given that the dataset is quite old (1988), there may have been changes in the prevalence of people with diabetes and in their risk factors. Since the data was collected however, medical knowledge as well as diagnostic criteria have changed, which could affect the relevance of some features or relationships in today's diabetes prediction.

Our prediction task is another limitation (diabetes vs. no diabetes) in that it is binary. Pre-diabetes is an important diagnostic category used in clinical practice, and incorporating this intermediate stage should therefore also be included in the prediction model in future studies.

Finally, although our deep learning models presented promising results, there are limitations on their interpretability. However, in the context of clinical applications, we need to be able to explain model decisions and further work in this line may entail developing more interpretable deep learning approaches to diabetes prediction.

Nevertheless, this study has produced insights that are useful for the domain of diabetes prediction in machine learning healthcare. This important medical task introduces promising performance capabilities for both traditional and deep learning approaches, and also emphasizes the requirement for additional research to overcome present limitations and enhance predictive accuracy.

VI. Conclusion

A. Summary of key findings

In our study of predicting diabetes using our machine learning and deep learning techniques, we have seen some exciting results. As a first step in this exploratory data analysis of the database of PIMA Indian diabetic data, we found that there are statistically significant relationships of significant health indicators with diabetes risk. Although correlation coefficients between glucose levels and diabetes outcome ($r = 0.47$) affirmed the key finding of blood sugar management in diabetes prevention and control, the wider applicability of the findings was questioned. Body mass index (BMI) and age were also found to have moderate positive correlations with diabetes outcome parameters thus suggesting several risk factors associated with diabetes.

We compared several traditional machine learning as well as deep learning as model approaches both of which yielded promising results in our predictive models. The traditional models, from best performing to worst, proved to be Logistic Regression with 75.32% accuracy, and Random

Forest lagging slightly with 74.61%. The results demonstrate that extracting meaningful predictive information from the provided features is possible with a relatively simple model.

And, as it turns out, our deep learning models, specifically, the Feedforward Neural Network (FNN) was better off, with an accuracy of 76.43%; There is the slight, but statistically significant, improvement over traditional models which hints at possibly complex, non linear relationships in the data that are more accurately reflected in more advanced algorithms. The adapted Convolutional Neural Network (CNN) presented very good results as well, showing this architecture to be fundamentally flexible to any structured data, and not restricted to image processing applications.

In addition, by analyzing our statistical results, we found that the FNN was statistically significantly outperformed by all of the traditional machine learning models, albeit no statistical difference was observed between the FNN and CNN. Therefore, these two deep learning biotechnologies can serve equally well at classifying diabetes in this dataset.

B. Implications for health-care

The results of this study are clinically important for healthcare in general, and for diabetes prevention and management in particular. Second, these methods have already proven that they can be good predictors, so, in particular machine learning models -- in particular, deep learning methods -- could be useful tools in supporting clinicians' decision systems. If healthcare providers can better identify individuals at high risk for developing diabetes, they can try to deliver intervention and prevention efforts more accurately to that population.

Our deep learning models are better than traditional statistical methods and simpler machine learning methods, suggesting that there may be some patterns in the subtle patterns, or very complicated relationships between the risk factors, that do not get caught by the current methods that work. This, again, proves that advanced analytics holds a lot of power in allowing us to better understand diabetes risk and progression.

In addition, our models provide us the means to determine key predictive features useful to us in clinical practice, in regards what health indicators we need to watch out for. For instance, the very close relationship between blood glucose levels and diabetes outcome makes continued blood glucose monitoring important in the undiagnosed, even if they are not yet diabetic.

Using a relatively small set of easily measurable health indicators that we demonstrate achieve high levels of accuracy, these models are especially promising for resource limited healthcare settings. It means that implementing diabetes risk assessment does not need any expensive medical tests or large amount of time.

While our models offer promising predictive power, the important point to reiterate is that our models should be viewed as being useful aids, but not substitutes, to clinical judgment. A health care system should be careful about integrating the predictive models into the health care systems; and the predictive models should be validated and refined in real world with the performance.

C. Future research directions

Our work has greatly pushed the use of machine learning to predict diabetes, and suggests lots of directions for future work. Because our study was conducted based on the PIMA Indian Diabetes dataset, which is specific for a particular population, and is also dated, further studies need to validate these models using larger, more diverse, or more recent datasets. This would facilitate the assessment of generalizability of our findings as well as identification of population specific risk factors.

Second, future research should also include longitudinal data. For our current study we used cross-sectional data but we thought it would be cool to better understand how diabetes risk grows and exploit how much additional value is obtained by predicting different features over longer time spans.

A second important direction for future work consists of building more interpretable deep learning model. While our deep learning models performed better, they are 'black box,' and we found they can act as a barrier to clinical adoption. Techniques such as SHAP (SHapley Additive exPlanations) values or LIME (Local Interpretable Model-agnostic Explanations) could suppore twger more transparent explanation of model predictions.

Additionally, future studies could incorporate other data types that are absent in the current dataset: both the genetic information or the lifestyle factors. This could potentially increase predictive accuracy and supply a better 'complete' picture of diabetes risk, they said.

Finally, future research should be devoted to the implementation of these predictive models into real world clinical settings. In this case, technical implementation of these tools would not be enough, but rather scientific studies on how these tools affect clinical decision making and patient's outcome.

Finally, due to diabetes often occurring along with many other chronic diseases in the future, we will study multi task learning methods which simultaneously learn to compute the risk for diabetes and the risk of these other diseases. It might give a better picture of someone's health state and risk profile.

We also demonstrate that machine learning and deep learning can be used to improve diabetes prediction. These methods potentially enhance diabetes related health outcomes for people at risk by early detection and management of diabetes.

REFERENCES

Mukhtar, Y., Galalain, A., & Yunusa, U. (2020). A modern overview on diabetes mellitus: a chronic endocrine disorder. European Journal of Biology, 5(2), 1-14.

World Health Organization. (2016). Global Report on Diabetes. World Health Organization.

Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., & Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Scientific reports, 10(1), 11981.

American Diabetes Association. (2021). Standards of medical care in diabetes—2021. Diabetes Care, 44(Supplement 1), S1-S232.

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics, 9, 515.

Kandhasamy, J. P., & Balamurali, S. (2015). Performance analysis of classifier models to predict diabetes mellitus. Procedia Computer Science, 47, 45-51.

Verma, V., Verma, S. K., Kumar, S., Agrawal, A., & Khan, R. A. (2024, July). 8 Diabetes Classification and Prediction Through Integrated SVM-GA. In Recent Advances in Computational Intelligence and Cyber Security: The International Conference on Computational Intelligence and Cyber Security (p. 96). CRC Press.

Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia Computer Science, 132, 1578-1585.

Teles, G., Rodrigues, J. J., Rabelo, R. A., & Kozlov, S. A. (2021). Comparative study of support vector machines and random forests machine learning algorithms on credit operation. Software: Practice and Experience, 51(12), 2492-2500.

Rupapara, V., Rustam, F., Ishaq, A., Lee, E., & Ashraf, I. (2023). Chi-square and PCA based feature selection for diabetes detection with ensemble classifier. Intell. Autom. Soft Comput, 36(2), 1931-1949.

Kasula, B. Y. (2023). Machine Learning Applications in Diabetic Healthcare: A Comprehensive Analysis and Predictive Modeling. International Numeric Journal of Machine Learning and Robots, 7(7).

Jayatilake, S. M. D. A. C., & Ganegoda, G. U. (2021). Involvement of machine learning tools in healthcare decision making. Journal of healthcare engineering, 2021(1), 6679512.

Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., & Malanda, B. (2018). IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Research and Clinical Practice, 138, 271-281.

Shehab, M., Abualigah, L., Shambour, Q., Abu-Hashem, M. A., Shambour, M. K. Y., Alsalibi, A. I., & Gandomi, A. H. (2022). Machine learning in medical applications: A review of state-of-the-art methods. Computers in Biology and Medicine, 145, 105458.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from

overfitting. The Journal of Machine Learning Research, 15(1), 1929-1958.

Zhu, T., Li, K., Herrero, P., & Georgiou, P. (2020). Deep learning for diabetes: a systematic review. IEEE Journal of Biomedical and Health Informatics, 25(7), 2744-2757.

Deo, R. C. (2015). Machine learning in medicine. Circulation, 132(20), 1920-1930.

Qi, X., Lu, Y., Shi, Y., Qi, H., & Ren, L. (2024). A deep neural network prediction method for diabetes based on Kendall's correlation coefficient and attention mechanism. Plos one, 19(7), e0306090.

Hounguè, P., & Bigirimana, A. G. (2022). Leveraging pima dataset to diabetes prediction: case study of deep neural network. Journal of Computer and Communications, 10(11), 15-28.

Qi, X., Lu, Y., Shi, Y., Qi, H., & Ren, L. (2024). A deep neural network prediction method for diabetes based on Kendall's correlation coefficient and attention mechanism. Plos one, 19(7), e0306090.

MOUSA, A., MUSTAFA, W., MARQAS, R. B., & MOHAMMED, S. H. (2023). A comparative study of diabetes detection using the Pima Indian diabetes database. Journal of Duhok University, 26(2), 277-288.

Maniruzzaman, M., Rahman, M. J., Al-MehediHasan, M., Suri, H. S., Abedin, M. M., El-Baz, A., & Suri, J. S. (2018). Accurate diabetes risk stratification using machine learning: role of missing value and outliers. Journal of Medical Systems, 42(5), 92.