Model Prediction on Diabetes in Females over 21 Years Old of Pima Indian Heritage Using Neuron Network

Zixuan Zhao, Shaoke Qi, Zhenshuo Xu, Tianze Bo

***Abstract*—The rates of diabetes have increased these days as a consequence of the prevalence of high-sugar and high-calorie foods. findings from the Nurses' Health Study, Nurses' Health Study 2, and the Health Professionals Follow-up Study demonstrate a significant relationship between the intake of sweetened beverages, red and processed meats, and the development of diabetes. In our study, we will dig deeper into the essence of this. Our study uses the most common body parameters, pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI, diabetes pedigree function, and age to improve the prediction accuracy of diabetes existence. By implementing the deep neuron networks, this study intended to enhance the accuracy of diabetes prediction in the Pima tribe and reduce the impact of race on diabetes prediction. The experiment conducted based on this specific group, using multiple standard and easy-to-collect features, can thus be easily generalized and contribute to predicting diabetes among the world population.**

***Index Terms*—Diabetes Prevalence, Pima Indian Tribe, Neural Networks in Epidemiology, Indigenous Health, Element Levels and Diabetes, Predictive Modeling, Diabetes Risk Factors, Machine Learning in Medical Diagnosis**

I. Introduction

IN modern society, the variety of foods available to satisfy people's appetites often leads to increased consumption of sugar- and calorie-rich ingredients, resulting in a rise in diabetes cases (see Fig. 3) [8]. Epidemiological studies like the Nurses' Health Study, Nurses' Health Study 2, and the Health Professionals Follow-up Study have demonstrated a significant link between the incidence of type 2 diabetes and diets high in sweetened beverages, red meat, and processed meat. The second study highlighted this association, emphasizing the need for targeted dietary interventions [2]. Moreover, the existence of diabetes are extends to even younger ages over time (see Fig. 1 and Fig. 2) [9].Our research builds on these findings by using neural network models to improve diabetes prediction, focusing on a dataset of women from the Pima tribe [3]. This group predominantly consumes a traditional diet of corn, beans, squash, and Indigenous meats such as venison, rabbit, and fish [5]. Given that the Pima diet includes known risk factors for diabetes, this serves as an ideal case study to examine how dietary models can enhance the accuracy of diabetes prediction and reduce racial bias.

We should first understand what diabetes is and why we want to indicate diabetes so that we can build the model to predict the existence of diabetes in females in the Pima Indian Tribe. Diabetes will be diagnosed when a person's blood sugar levels are consistently high, which usually happens when the human body cannot successfully control the sugar level. One of the main contributors to blood sugar is glucose, which comes from the daily consumption of foods. Glucose is transported through the bloodstream so that each cell can access it. To safely utilize glucose, insulin, a specific hormone, plays a crucial role during transportation. The lack of this particular hormone can accumulate glucose in the bloodstream, creating hyperglycemia in the veins. If not corrected, persistent hyperglycemia can develop into diabetes, which poses a significant threat to a person's health [4].

Diabetes has many categories that have different causes and consequences. The main types are type 1 diabetes, type 2 diabetes, prediabetes, and gestational diabetes. Type 1 diabetes is an autoimmune disease in which the body's immune sys-

Characteristic	Diagnosed diabetes Percentage (95% CI)	Undiagnosed diabetes Percentage (95% CI)	Total diabetes Percentage (95% CI)
Total	**11.3 (10.3–12.5)**	**3.4 (2.7–4.2)**	**14.7 (13.2–16.4)**
Age in years			
18–44	3.0 (2.4–3.7)	1.9 (1.3–2.7)	4.8 (4.0–5.9)
45–64	14.5 (12.2–17.0)	4.5 (3.3–6.0)	18.9 (16.1–22.1)
≥65	24.4 (22.1–27.0)	4.7 (3.0–7.4)	29.2 (26.4–32.1)
Sex			
Men	12.6 (11.1–14.3)	2.8 (2.0–3.9)	15.4 (13.5–17.5)
Women	10.2 (8.8–11.7)	3.9 (2.7–5.5)	14.1 (11.8–16.7)
Race-Ethnicity			
White, non-Hispanic	11.0 (9.4–12.8)	2.7 (1.7–4.2)	13.6 (11.4–16.2)
Black, non-Hispanic	12.7 (10.7–15.0)	4.7 (3.3–6.5)	17.4 (15.2–19.8)
Asian, non-Hispanic	11.3 (9.7–13.1)	5.4 (3.5–8.3)	16.7 (14.0–19.8)
Hispanic	11.1 (9.5–13.0)	4.4 (3.3–5.8)	15.5 (13.8–17.3)

Fig. 1. 2017-2020 Diabetes Data

Characteristic	Diagnosed diabetes Number in Millions (95% CI)	Undiagnosed diabetes Number in Millions (95% CI)	Total diabetes Number in Millions (95% CI)
Total	**29.4 (26.7–32.0)**	**8.7 (7.0–10.5)**	**38.1 (34.2–42.0)**
Age in years			
18–44	3.5 (2.8–4.2)	2.2 (1.5–3.0)	5.8 (4.7–6.8)
45–64	12.0 (10.1–13.9)	3.8 (2.7–4.8)	15.8 (13.4–18.2)
≥65	13.8 (12.5–15.1)	2.7 (1.6–3.8)	16.5 (15.0–18.1)
Sex			
Men	16.1 (14.1–18.0)	3.7 (2.6–4.8)	19.8 (17.4–22.1)
Women	13.3 (11.5–15.1)	5.0 (3.3–6.7)	18.3 (15.3–21.3)
Race-Ethnicity			
White, non-Hispanic	17.8 (15.2–20.4)	4.3 (2.4–6.1)	22.1 (18.5–25.7)
Black, non-Hispanic	4.0 (3.3–4.6)	1.4 (1.0–1.9)	5.4 (4.7–6.1)
Asian, non-Hispanic	1.8 (1.5–2.1)	0.9 (0.5–1.2)	2.7 (2.2–3.1)
Hispanic	5.0 (4.3–5.7)	1.9 (1.4–2.4)	6.9 (6.2–7.6)

Fig. 2. 2021 Diabetes data

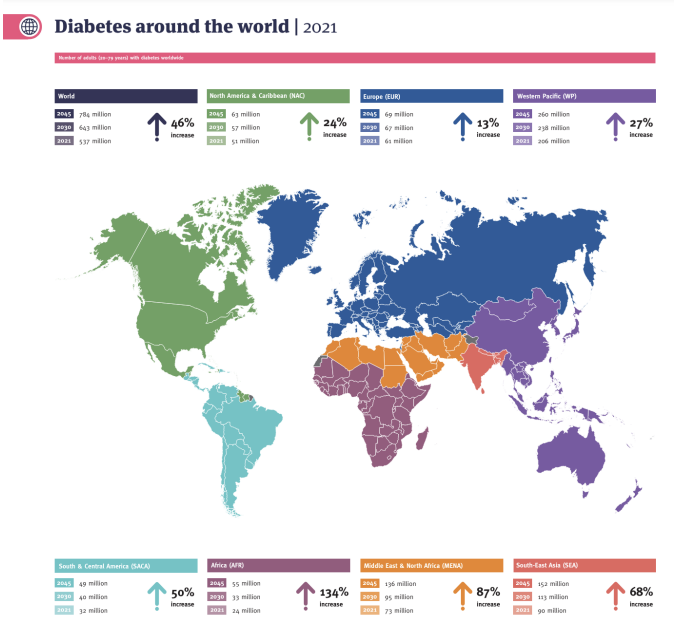

Fig. 3. 2021 Diabetes Trends over The World

tem attacks insulin-producing cells, leading to an insufficient supply of insulin. This type is less common in the general diabetes population. Prediabetes is a condition in which blood sugar levels are higher than average but not high enough to be classified as diabetes and is a precursor to type 2 diabetes. Type 2 diabetes is the most common form of diabetes worldwide. It occurs when the body does not produce enough insulin or does not use the insulin it produces properly. It usually occurs in adults and rarely in children. Gestational diabetes occurs during pregnancy and usually disappears after birth but increases the risk of developing type 2 diabetes later in life [4].

The consequences of diabetes are also various due to different types of diabetes as well as other body conditions. Persistently high blood sugar levels, the hallmark of type 2 diabetes, can lead to chronic dehydration and mental confusion. These symptoms not only reduce quality of life but, in severe cases, can be fatal and require immediate medical intervention. When the body has too little insulin, it cannot use glucose for energy. Instead, it begins to metabolize fats, a process that produces ketones. An accumulation of ketones in the blood can lead to a dangerous condition called ketoacidosis, which is characterized by symptoms such as unconsciousness, vomiting, and breathing difficulties and is often associated with type 1 diabetes. In addition, diabetes causes several other serious health problems, including strokes and heart attacks, which seriously affect overall health [4].

We utilize a dataset with eight contributors to predict the existence of diabetes in a specific person in the Pima Indian Heritage[3]. The eight factors, pregnancy, blood sugar level, blood pressure, skin thickness, insulin level, body mass, diabetes pedigree function, and age, are used together to construct the final model. Since the output is binary (1 or 0), we decided between the perceptron model and the Deep Neuron Network. Since the Pima Indian tribe follows dietary habits corresponding to the high risk of diabetes, this population will be a good choice in constructing our prediction model. By focusing on this specific population, we hope to develop a predictive tool that will not only deepen our understanding of diabetes in this group but also contribute to the broader prevention and treatment of diabetes through interventions tailored to the community's specific needs. This article is organized as follows: review of previous studies, exploratory data analysis, methodological discussion, evaluation of training procedures and model testing, conclusions and discussion of the prediction model, and a reference page.

II. Literature Review

Recent advances in machine learning (ML) have greatly improved the ability to predict diabetes, an urgent global health problem. Several studies have used different ML models to improve the accuracy of diabetes prediction in other populations. This literature review presents three pioneering articles on predicting diabetes using machine learning. These studies are compared with this study's unique approach, which uses neural networks to predict diabetes in women of the Pima tribe in India.

The research paper by Tasin et al. *Predicting Diabetes Using Machine Learning and Explainable Artificial Intelligence Techniques* examines explainable artificial intelligence (AI) techniques in a population of Bengali women to predict diabetes using machine-learning models such as XGBoost. The study's strength is the integration of ADASYN and LIME to handle unbalanced data and ensure the interpretability of the models. However, the study primarily focused on a homogeneous population in terms of race and geographic location [6].

A review of diabetes risk prediction using machine-learning approaches by Firdous et al. examines the various machine-learning methods used to predict diabetes risk. Several strategies are explicitly used in the early prediction of diabetes, including SVM, KNN, and Random Forests. Despite the extensive analysis, the study leaned heavily on general machine learning algorithms. It did not address the influence of specific diets or racial and ethnic nuances on the prevalence of diabetes [7].

Ahmed et al.'s research on diabetes prediction based on machine learning and the development of intelligent Web applications focused on developing a Web application for diabetes prediction using the tool Flask, which applies a set of machine learning algorithms to clinical data. Although this study innovatively addressed the development of user-friendly applications for diabetes prediction tools, it did not explicitly address the influence of nutrition or the unique genetic and environmental factors associated with specific ethnic groups[1].

In contrast, this study attempts to fill these gaps by focusing specifically on the Pima-Indian population, the population that has not only a high appearance of diabetes but also a high-diabetes-risk daily diet. Unlike the studies mentioned above, this study synthesizes epidemiological findings from extensive

health studies, such as the Nurses' Health Study and the Health Professionals Follow-Up Study, that have associated specific dietary patterns with diabetes risk. Combining all these body parameters into machine learning algorithms, this paper is intended to provide a prediction model that can account for racial factors that are usually ignored during the diagnosis of diabetes.

III. Exploratory Data Analysis

A. Data Description

We utilize the dataset collected on the females of Pima Indian Heritage [3]. The dataset has 768 rows and nine columns, indicating the 768 specific individuals with their targeted body parameters. Using the standard provided by the World Health Organization (WHO), which was 126mg/dl, the dataset determines whether the specific individual has diabetes.

B. Variables

The dataset contains a total of nine variables described as follows:

- **Pregnancies**: Number of times the patient has been pregnant.
- **Glucose**: Plasma glucose concentration 2 hours in an oral glucose tolerance test (mg/dL).
- **Blood pressure**: Diastolic blood pressure (mm Hg).
- **Skin Thickness**: Triceps skin fold thickness (mm).
- **Insulin**: Two-hour serum insulin (mu U/ml).
- **BMI**: Body mass index (weight in kg/(height in m)$\hat{2}$).
- **Diabetes Pedigree Function**: A function that scores the likelihood of diabetes based on family history.
- **Age**: The patient's age in years.
- **Outcome**: A binary variable indicating the presence (1) or absence (0) of diabetes.

C. Statistical Data Summary

- The **Glucose**, **Blood Pressure**, **Skin Thickness**, **Insulin**, and **BMI** variables initially contained zero values, which were identified as placeholders for missing data. However, to ensure the prediction result corresponds to the observed data, we decided to treat these zero values as extreme values so that they can still be considered for the final prediction.
- The **Glucose** levels ranged from 44 to 199 mg/dL with a mean of approximately 121 mg/dL, indicating a wide spread of glucose concentrations (see Fig. 5).
- **Blood Pressure** readings varied from 24 to 122 mm Hg, with a mean value of 72 mm Hg, suggesting normal blood pressure variation among the participants (see Fig. 4).
- **Skin Thickness** measurements showed significant variability, ranging from 7 to 99 mm (see Fig. 6).
- **Insulin** levels were highly skewed, ranging from 14 to 846 mu U/ml, reflecting the diverse insulin therapy regimes or endogenous insulin production among subjects (see Fig. 5).
- The **BMI** ranged from 18.2 to 67.1, with a considerable number of individuals categorized as overweight or obese, a common risk factor for diabetes (see Fig. 4).

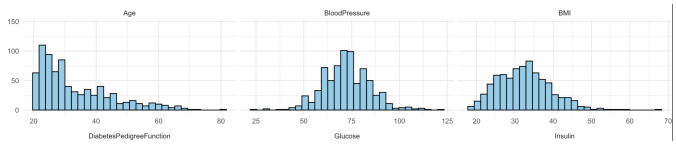

Fig. 4. Distribution of Age, Blood Pressure, and BMI

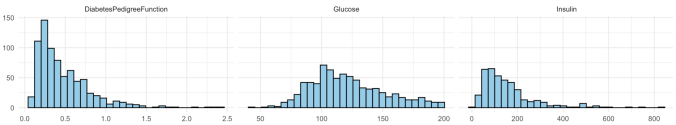

Fig. 5. Distribution of Diabetes Pedigree Function, Glucose, and Insulin

- The overall distributions for all observed data are shown in Fig. 4,5, and 6.

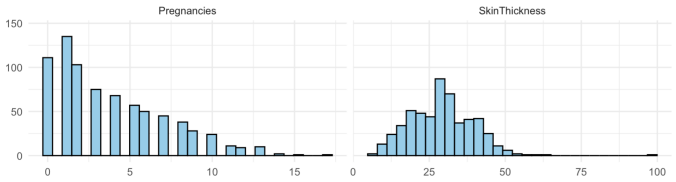

Fig. 6. Distribution of Pregnancies and Skin Thickness

D. Box Plot

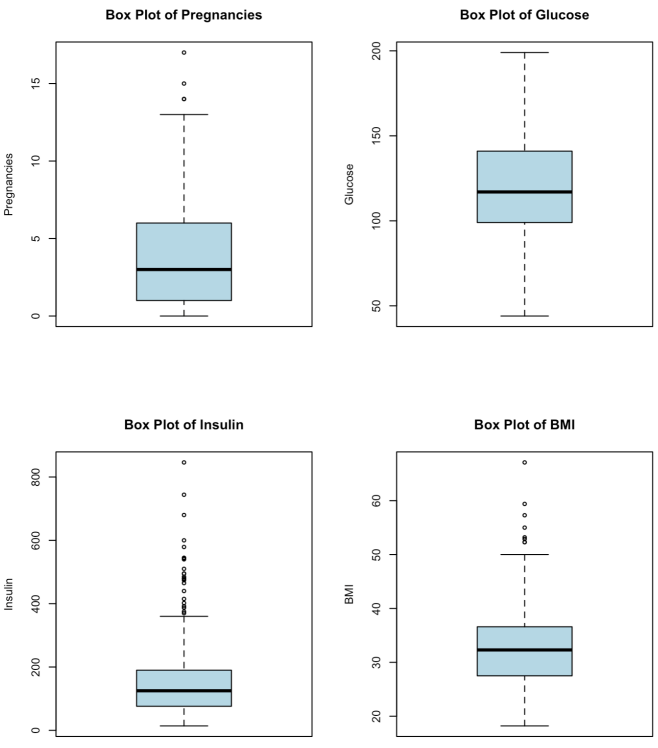

Fig. 7. Box Plots for Pregnancies, Glucose, Insulin, and BMI

We construct box plots (see Fig. 7 and 8) for each individual to understand the distribution within each body parameter

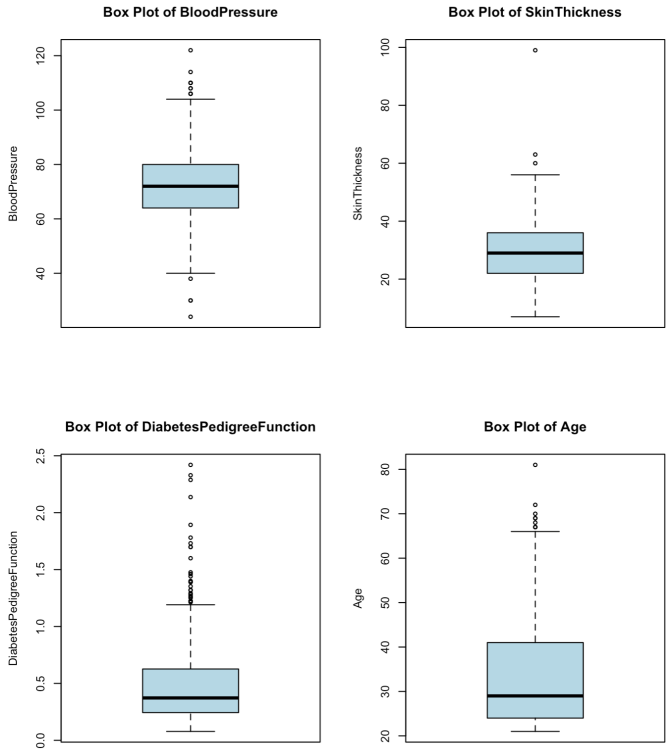

Fig. 8. Box plot for Blood Pressure, Skin Thickness, Diabetes Pedigree Function, and Ages

we are interested in. Such a boxplot can explicitly explain the targeted parameters' median and variability, including pregnancy, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age.

The glucose boxplot shows the median, 110mg/dl, lying in the densest area of the whole distribution of glucose, the range of 100 to 140 mg/dl (see Fig. 7). Based on its distribution pattern and small number of outliers, we can conclude that this variable is relatively stable. In contrast, the blood pressure boxplot shows a different pattern. With the median lying at 70 mmHg, the interquartile blood pressure range is significantly smaller, creating many outliers from the observed data (see Fig. 8).

The box plot for insulin level provides the most significant meaning during the analysis. With a low median and a small interquartile range, the insulin level data distribution shows a severe pattern of outliers (see Fig. 7). This pattern indicates that insulin level can significantly contribute to the prediction model when constructing the model. Following the same pattern, Skin Thickness is also an essential contributor to building a model, as it has a relatively low median with a narrow interquartile range, showing a significant pattern of outliers. (see Fig. 8).

BMI, with a significant distribution of observed data between 25 and 50, shows that obesity and being overweight are prevalent in the observed data (see Fig. 7). These two factors are considered major contributors to diabetes. Therefore, this should also be a significant parameter to consider. The diabetes pedigree function, a measure of genetic risk factors, showed a wide range of values and some extremes, indicating different genetic predispositions in the population (see Figure 5).

Lastly, the box plot of age indicates that the observed data mainly consists of middle-aged people, the majority of people with diabetes. Therefore, this can also be a significant index to be considered during the model construction (see Fig. 8).

These box plots are essential for determining the range, central tendency, and variance of each variable, as well as for identifying potential outliers or anomalies in the data that may require further investigation or specialized analysis methods during the modeling phase of the study.

E. Data Quality Analysis and Handling Methods

Many data need to be cleaned because of biologically improbable values, such as zeros for physiological measurements that should not be zero (e.g., blood pressure and body mass index). These values were treated as outliers to ensure that the results of training and testing our model can be used for prediction without loss of generalizability, as anomalous data are expected during research.

The cleaned dataset, therefore, provides a reliable basis for investigating the relationship between physiological measurements and diabetes status and emphasizes the need to generate statistically valid and clinically relevant results.

F. Advance Data Visualization

After initially addressing data quality issues such as zero values, the following steps involve more profound visualizations and analyses to explore the relationships within the data:

1) Correlation Matrix: We created a correlation matrix (see Figure 9) to understand the relationship between variables and their potential influence on diabetes outcomes. We can explicitly interpret the relationships between any pairs of variables from the observed body parameters. Among the forms of the correlation matrix, a heat map should be considered the best choice for this scenario since we have a significant amount of data that needs a most sensitive reflector, typically color, to visualize the result.

This matrix helps identify relationships that may affect the diabetes outcome prediction model.

In the heat map, each small block represents the correlation index, ranging from -1 to 1, reflected by its lightness of color. Values close to -1 and 1 indicate a strong negative or positive correlation between the two data, meaning they have significant interaction effects. In contrast, a value close to zero indicates loose relationships between the two parameters.

Notably, several variables show significant correlations with each other and with the diabetes outcome:

- **Glucose and Outcome**: The correlation coefficient of 0.47 between glucose levels and diabetes outcome underscores the strong relationship between high glucose levels and the likelihood of diabetes, which is well-documented in diabetes research.
- **BMI and Insulin**: BMI and insulin levels show a moderate positive correlation of 0.2. This suggests that higher body mass index, a marker of obesity, is associated

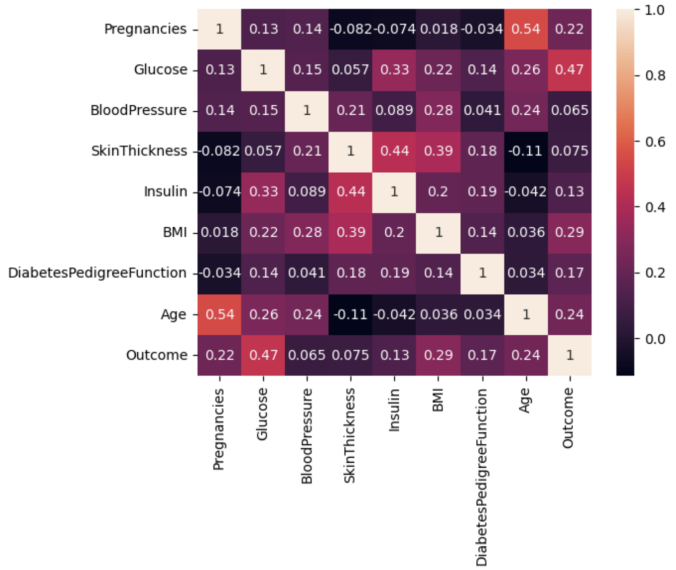

Fig. 9. Correlation Matrix for All Observed Predictors Data

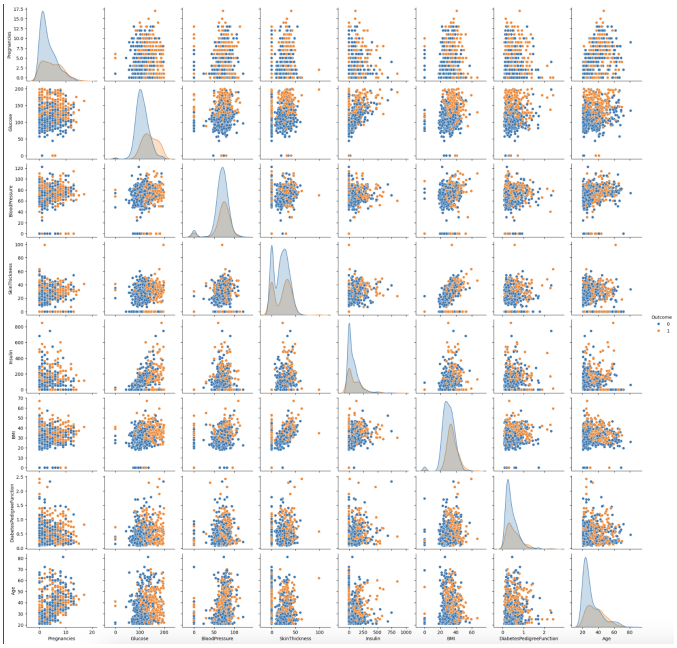

Fig. 10. Pair plots for All Observed Predictors Data

with increased insulin levels, possibly reflecting insulin resistance.

- **Skin Thickness and Insulin**: Skin thickness and insulin correlate 0.44. This relationship might indicate that higher skin fold thickness, a surrogate marker of body fat, is associated with greater insulin levels, further suggesting insulin resistance.

Other notable observations include:

- **Blood Pressure and Age**: Blood pressure shows a modest correlation with age (0.24), indicating that blood pressure tends to increase with age.
- **Diabetes Pedigree Function**: This genetic risk factor shows only weak correlations with other variables and the outcome, suggesting that while it has a role in diabetes risk, it does not strongly influence other physiological measures in this dataset.

Overall, the correlation matrix provides critical insights into the inter dependencies among variables, highlighting those that may have significant roles in the development of diabetes. These findings are essential for selecting features in developing a predictive model, as they help focus on variables with the most significant potential predictive power for diabetes. Understanding these relationships also aids in crafting more nuanced intervention strategies, potentially improving diabetes prevention and management in populations similar to the one studied here.

2) Pair Plots: We constructed a pair plot based on scaled value (see Fig. 10) to investigate the bivariate relationships between key variables. These plots provide a comprehensive view of how different variables relate to each other through scatter plots for each pair of variables, while histograms on the diagonal allow for an examination of the distribution of each variable. Such detailed visualization aids in spotting trends, patterns, or anomalies that might not be evident from the raw data or statistical summaries alone. The pair plots are significant in identifying interaction features and propose the unreducible model predictors.

Each plot represents one variable's interaction with the other, showing their combined contribution to the output value. The orange point represents the existence of diabetes, and the blue points indicate there is not. The orange points in the top right corner indicate that the two interacted factors positively correlate with the output. In contrast, the blue point aligned at the top right corner indicates the opposite, which is the non-existence of diabetes. The pair plots directly reflect the interaction contribution of each pair of factors to the outcome.

For the pair plot, we intended to find out the interaction plots that indicate a high risk of diabetes when the two-parameter both have higher values (visually represented by the clustering of the orange point at the top right corner):

- **Glucose and Pregnancies onto Outcome:** The higher the glucose level and the more times pregnancies are formed together, the higher the risk of having diabetes. This phenomenon is explained by the type of diabetes sections, which demonstrate that pregnancy is a potential factor of diabetes and glucose is the main contributor to diabetes.
- **BMI and Glucose onto Outcome:** A higher BMI indicates higher body mass relative to height indexes, which implies a person's fatness. The higher the BMI is, the more unhealthy a person can be. With a higher glucose level, a person with a higher BMI can be exposed to a higher risk of having diabetes.
- **Glucose and Age onto Outcome**: As indicated in the research in the introduction, we can see that the older a person can be, the weaker the body will be. As the glucose level increases, the risk of having diabetes increases significantly.
- **BMI and Age onto Outcome**: The weaker body with

an unhealthy condition is weak to any disease that significantly impacts a person. This can be intuitively interpreted through the data.

Overall, the pair plots provide critical insights into the interactions among variables and the outcome, highlighting those factors that should be considered together. They give us a fundamental prediction of the final model predictors. Understanding these interactions also provides solid evidence for the formation of diabetes.

IV. METHODOLOGY

First, we pre-processed the diabetes.csv dataset. After that, we split the dataset into two parts: the large part for training and the small part for testing. We used the training data to train two models: Perceptron and Deep Neural Network. Then, we evaluated these models' performance by scores which computed in different metrics. Finally, the best model is deployed in a web application using flask to illustrate our work.Following this, we describe the workflow of each part briefly: - 1 Data Collection: Our project adopts the diabetes.csv dataset as our training set. This dataset contains eight features for diabetes prediction and was carefully chosen to ensure a solid basis for model construction and assessment. - 2 Data Analysis and Data Preprocessing: In our study, we applied data standardization to minimize the negative effect of the value of data. Additionally, to achieve the ablation study, we constructed eight new sub-datasets by removing one of the eight features in each dataset. This approach allowed us to analyze the how much impact each feature has on the models' performance. - 3 Model Construction and Prediction: To construct the model, 80% of the diabetes data is used for training while 20% of diabetes data is used for the testing validation. - 4 Performance Analysis: We have analyzed the results of our model in terms of multiple performance metrics. The algorithm that provides the highest prediction accuracy is selected as the best algorithm for web application illustration.

A. Adopted machine learning algorithms

In this section, we will describe various machine-learning algorithms that are used in the predictive model.

•Perceptron: The Perceptron is a type of artificial neuron used in machine learning and artificial intelligence. It is a combination of multiple inputs and applies a step function and a single output. The perceptron algorithm updates the weights derived from the prediction error, aiming to find the optimal weights to correctly classify the data. The model can be described as:

$$y = \begin{cases} 1 & \text{if } (w \cdot x + b) \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Here, w is the weight vector, x is the input vector, and b is the bias.

•Deep Neural Network: A Deep Neural Network consists of multiple layers of artificial neurons, including input layers, hidden layers, and output layers. Each neuron receives input from the previous layer, processes it, and passes the result to the next layer. The output of a neuron in a layer can be described as:

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right)$$

Here, f is the activation function (e.g., ReLU, sigmoid), w_i are the weights, x_i are the inputs, and b is the bias. Training a DNN involves backpropagation and optimization techniques to minimize the error and improve the model's accuracy.

V. EVALUATION

To evaluate the performance of our proposed models, we conducted extensive experiments on the diabetes.csv dataset. The loss function used was binary cross-entropy loss. Our network was trained using SGD for 100 epochs with the tensorflow library. We set the learning rate to 0.01.

A. Performance Metrics

The performance of the machine learning models was evaluated using several metrics, including accuracy, precision, recall, and F1-score. These metrics are defined as follows:

1) Accuracy: It measures the model's total number of accurate predictions and can be measured as a ratio between the number of correct. Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$

2) Precision: The proportion of correct positive predictions to total positive predictions is known as precision. Precision $= \frac{TP}{TP+FP}$

3) Recall: Recall: Total positive predictions vs. actual positive values are known as recall. Recall $= \frac{TP}{TP+FN}$

4) F1-score: F1-score takes precision and recall into account and can be described as follows. F1-score $= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

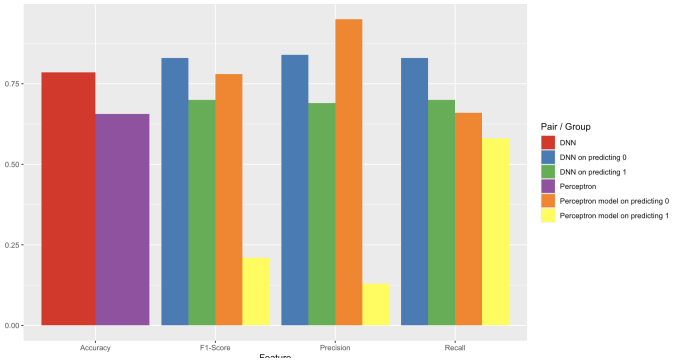

Fig. 11. Comparison of Perceptron and DNN Performance

B. Results of diabetes dataset

Pima Indian Diabetes Dataset [3] is one of the ideal datasets for evaluating machine learning algorithms for predicting diabetes. The National Institute of Diabetes and Digestive and Kidney Diseases provided the Pima Indian dataset [3] to determine if a patient has diabetes based on diagnostic measures like Pregnancies, Glucose level, Blood Pressure, Skin Thickness, Diabetes Pedigree Function, Insulin, BMI and Age . Fig. 11 represents the overall results of the experiment in

Removed Feature Name	Perceptron Learning	DNN
Pregnancies	53.89%	75.97%
Glucose	55.84%	66.88%
BloodPressure	65.58%	77.27%
SkinThickness	65.58%	74.03
Insulin	70.78%	77.92%
BMI	60.39%	76.62%
DiabetesPedigreeFunction	58.44%	75.32%
Age	59.09%	79.87%
/	65.58%	78.58%

TABLE I
PERCEPTRON AND DNN MODEL ACCURACY SCORE COMPARISON

terms of accuracy, precision, recall and f1-score. The accuracy of the various models is 65.58 Percent and 78.57 Percent for Perceptron, Deep Neural Network, respectively. It also indicates that DNN provides better performance.

In next experiment, we find most influential attributes using correlation matrix that states how the features are related to each other on the target variable. Figure 6 shows the correlation matrix between each of the attributes to the class variable. The relationship between the parameters is depicted in the correlation plot. The most associated parameters with the Outcome are glucose, age, BMI, and pregnancies. • Insulin and Diabetes Pedigree Function have no bearing on the final result. • There is only a slight connection between blood pressure and skin thickness and the outcome. • Age and Pregnancy, Insulin and Skin Thickness, BMI and Skin Thickness, Insulin and Glucose all have a small association.

C. Ablation Study

We implement extensive ablation experiments on the diabetes dataset to inspect the significance of each feature in our model. All results are precision and f1-score on eight control datasets from the original diabetes dataset.

We constructed eight subsets of the data, each removing one of the features, and evaluated the models on these modified datasets. We computed the accuracy and F1-score for each subset to understand the impact of each feature on the model's performance (see Table 1). The dataset with the Skin Thickness feature removed had the highest performance, while the dataset with the glucose level feature removed had the worst performance. This suggests that glucose level plays a critical role in our predictive model, whereas the Skin Thickness feature might introduce noise or irrelevant information that hampers the model's predictive capability.

While Skin Thickness is traditionally considered a significant factor in diabetes risk, its exclusion improving model performance indicates that in this specific dataset, Skin Thickness may not align well with the other features or may be correlated with other included variables, leading to redundancy or multicollinearity. This finding suggests that the presence of Skin Thickness may have been more detrimental than beneficial to the model's accuracy in this context.

These results highlight the necessity for a detailed examination of feature interactions within the dataset. While Skin Thickness is clearly vital for accurate predictions, the negative impact of Skin Thickness's inclusion points to a complex relationship between features that warrants further investigation. Understanding these interactions can lead to more effective data preprocessing and feature selection strategies, ultimately improving the model's performance.

VI. CONCLUSION

Our panel study was on female diabetic females over 21 years of age in Pima Indian Heritage [3]. Based on their dataset, we learn and test the model to determine more accurately whether a patient is experiencing diabetes. First, we understand that diabetes is a chronic disease that affects quality of life. Specifically, diabetes is a disease in which there is too much blood sugar in the body, which comes from the food we eat every day. These foods include carbohydrates and glucose. Insulin problems lead to insufficient insulin and high blood sugar. The other condition is diabetes due to insulin resistance. In this study, we developed and evaluated a neural network-based diabetes prediction model.

Through testing, we found that the neural network is better than the linear and perceptron models. Specifically, after data preprocessing, the accuracy rate of the neural network model reached 78%, the F1 score was 0.83, and the AUC was 0.79, showing good prediction ability.

Looking ahead, we are optimistic about the potential of this study. By expanding our data collection and patient population, we believe we can significantly enhance the prediction accuracy and generalization ability of our model. This research has not only deepened our understanding of machine learning and data preprocessing techniques but also underscored the critical role of data quality and model selection in predictive systems.

VII. AUTHOR CONTRIBUTION

Zixuan Zhao: Data Selection, Introduction, Literature Review, Exploratory Data Analysis, Latex Paper Edition. Shaoke Qi: Coding, Model Building and Model Training, Conclusion. Zhenshuo Xu: Web-based Front End, Model Testing. Tianze Bo: Evaluation, Code Commenting.

WORKS CITED

[1] Nazin Ahmed, et al., "Machine learning based diabetes prediction and development of Smart Web Application," *International Journal of Cognitive Computing in Engineering*, vol. 2, June 2021, pp. 229–241, https://doi.org/10.1016/j.ijcce.2021.12.001.

[2] Andres V. Ardisson Korat, et al., "Diet, Lifestyle, and Genetic Risk Factors for Type 2 Diabetes: A Review from the Nurses' Health Study, Nurses' Health Study 2, and Health Professionals' Follow-up Study," *Current Nutrition Reports*, U.S. National Library of Medicine, 1 Dec. 2014, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4295827/#:~:text=Higher%20consumption%20of%20coffee%2C%20whole,is%20associated%20with%20increased%20risk.

[3] UCI Machine Learning, "Pima Indians Diabetes Database," *Kaggle*, 6 Oct. 2016, https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?resource=download.

[4] Cleveland Clinic medical professional, "Diabetes," *Cleveland Clinic*, https://my.clevelandclinic.org/health/diseases/7104-diabetes. Accessed 5 June 2024.

[5] Laura Redish and Orrin Lewis, "Pima Indian Fact Sheet," *Facts for Kids: Pima Indians (Akimel O'odham, Pimas)*, https://www.bigorrin.org/pima_kids.htm. Accessed 5 June 2024.

[6] Isfafuzzaman Tasin, et al., "Diabetes prediction using machine learning and explainable AI techniques," *Healthcare Technology Letters*, vol. 10, no. 1–2, 14 Dec. 2022, pp. 1–10, https://doi.org/10.1049/htl2.12039.

[7] Gowher A. Wagai, et al., "A survey on diabetes risk prediction using machine learning approaches," *Journal of Family Medicine and Primary Care*, vol. 11, no. 11, 2022, p. 6929, https://doi.org/10.4103/jfmpc.jfmpc_502_22.
[8] "Factsheets." IDF Diabetes Atlas, https://diabetesatlas.org/regional-factsheets/?dlmodal=active&dlsrc=https%3A%2F%2Fdiabetesatlas.org%2Fidfawp%2Fresource-files%2F2021%2F11%2FIDFDA10-global-fact-sheet.pdf. Accessed 5 June 2024.
[9] "National Diabetes Statistics Report." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, https://www.cdc.gov/diabetes/php/data-research/index.html. Accessed 5 June 2024.