Machine Learning Approach in Medical Diagnosis: Predicting Diabetes Complications

Ling Jun Yuan
School of Computing
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumpur, Malaysia
TP065280@mail.apu.edu.my

Than Chi Ren
School of Computing
Asia Pacidic University of Technology and Innovation (APU)
Kuala Lumpur, Malaysia
TP065236@mail.apu.edu.my

Ch’ng Khai Nian
School of Computing
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumpur, Malaysia
TP064872@mail.apu.edu.my

Dr. Kamalanathan Shanmugam
Senior Lecturer / School of Technology
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumput, Malaysia
kamalanathan@apu.edu.my

Dr.Adeline Sneha J
Senior Lecturer / School of Computing
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumput, Malaysia
adeline.john@apu.edu.my

Juhairi Aris Muhamad Shuhili
Lecturer / School of Computing
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumpur, Malaysia
juhairi.shuhili@apu.edu.my

Abstract— **Nowadays, diabetes has become a worldwide disease that will have a bad impact on the human body’s health. Understanding body condition and detecting the signs of diabetes early is key to preventing diabetes from becoming serious. Therefore, an effective machine learning technology is implemented for predicting diabetes with different features from the diabetes dataset. This research is aimed to implement the machine learning model (Logistic Regression) to predict diabetes and identify the effect of parameters on accuracy. The real diabetes patient data set with 390 records is from Data World and applied to the model. The parameter which is “penalty” is set with different values to test the accuracy of the model. 30% of data will be used for testing data and 70% of data as training data. The results of model accuracy are more than 90%. The model is implemented well to predict diabetes and further experiments for testing the model are needed to improve the accuracy of the model.**

Keywords— Machine Learning, Logistic Regression, Medical Diagnosis, Diabetes, Diabetes Prediction

I. Introduction

The newest record of worldwide diabetes cases reported by **Institute for Health Metrics and Evaluation** (Institute for Health Metrics and Evaluation, n.d.) achieved more than half a billion people are having diabetes and the number of cases is estimated to more than double to 1.3 billion people in next 30 years. Diabetes disease with high blood sugar may have a big negative impact on many parts and organs, such as the legs, kidneys, eyes, heart, and feet, and it is possible to get a heart attack when getting severely obese. Diabetes may also cause high blood pressure and blood vessel obstruction. Thus, understanding personal obesity issues in advance and getting treatment earlier is indispensable in daily life to build a healthy body and happy life.

Machine learning (ML) technology keeps growing and expanding in many aspects. ML is about training the algorithms and models to let the machine learn without programming, using different algorithms and models to find valuable patterns and extract meaningful knowledge from a large amount of data. Medical diagnosis is one of the areas where ML techniques are applied to make the diagnosis and prediction of disease more advanced, accurate, and efficient. To identify and diagnose diseases like Alzheimer's disease, diabetes, breast cancer, heart disease, etc., numerous machine-learning algorithms have been developed and used by researchers. There are several techniques developed by researchers and used in medical diagnosis such as logistic regression, linear regression, Naïve Bayes, decision tree, K-nearest neighbors (KNN), random forest, and support vector machine (SVM). Because of ML approaches, predicting and discovering diabetes at the early stage is workable and the other illnesses from diabetes can be avoided. Lastly, the signs of obesity can be detected early, and the patients can make lifestyle changes to avoid diabetes.

II. Literature Review

A. *Application of machine learning in different medical diagnosis*

There are several effective machine learning approaches that are used by researchers to do the prediction of breast cancer. In the research of Khalid et al. (2023) shown that the different algorithms had been used to diagnose the breast cancer such as decision tree (DT), random forest (RF), logistic regression (LR), k-nearest neighbors (KNN), linear support vector classifier (linear SVC) and support vector classifier (SVC). The exploratory data analysis (EDA) dataset was getting from Kaggle.com and applied to the algorithms. From the experiments they conducted, RF had the highest accuracy with 96.49% followed by DT (93.86%), LR (92.98%), KNN (92.11%), linear SVC (89.47%) and SVC with 87.72% accuracy.

Rabiei (2022) demonstrated different machine learning models which are Genetic Algorithm (GA), Multi-layer Perceptron (MLP), Gradient Boosting trees (GBT), Random Forest (RF) and the data set contains 5178 records of people was applied in the models. Adaptive heuristics or search engine algorithms, genetic algorithms are mostly used in machine learning to tackle search and optimization problems. It is a method that addresses limited and unconstrained optimization problems by means of natural selection. In the result shown that RF model achieved 80% accuracy, GBT achieved 74% accuracy and MLP achieved 73% accuracy.

Efficient Heart Disease Prediction System is a decision tree algorithm technique to predict heart disease (Purushottam et al. 2016). Different data sets related to coronary disease were used to train and test the system. The important parameters of the model were set with specific values which are Confidence value was 0.25, MinItemsets was 2, and Threshold was 10. *Classified Rules*, *Rules without duplicates, Pruned Rules,* and *Original Rules* these four rules were generated to the experiment. The performance evaluation for their system achieved 86.7% accuracy.

Bhatt et al. (2023) applied different algorithms to detect heart disease and assessed their performance. The algorithms including multilayer perceptron (MLP), XGBoost, decision tree (DT) and random forest (RF) were used to train and test their performance. The dataset which includes 70,000 patient records and 12 different variables were applied in the models. The clustering method was applied to the dataset to group the instances based on the similarity measures. DT model was achieved 86.53% accuracy, 86.92% accuracy was achieved by RF, XGBoost was achieved 87.02% accuracy with the value 0.1 for parameters *'learning_rate'*, value of 4 for *'max_depth'*, value of 100 for *'n_estimators'*, value of 10 folds for *'cross-validation'* including 21,000 testing and 49,000 training data instances and finally MLP was achieved the highest accuracy with 86.94% after hyperparameter tuning.

Support vector machine (SVM) algorithm was used by Neelaveni & Devasana (2020) to detect the Alzheimer's disease. R programming, e1071 packages in R, some formula, and psychological parameters were being used to train the SVM algorithm. Their model was getting the accuracy of 85% in detection of Alzheimer's patients.

Zhang et al. (2023) experimented to classify Alzheimer's disease by using SVMs and random forest (RF). The Alzheimer's Disease Neuroimaging Initiative (ADNI) was used and applied to the SVM algorithm, ADNI and AddNeuroMed datasets were used to test the random forest algorithm. RF was achieved 86.6% accuracy for ADNI datasets and 86.25% for AddNeuroMed datasets while SVM achieved 88.48% accuracy when testing the ADNI datasets.

Dashtipour et al. (2022) proposed their algorithms and models to detect Alzheimer's disease including Naïve Bayes, logistic regression, k-nearest neighbors (KNN), support vector machine (SVM), decision tree, and random forest. The radial basis function (RBF) kernel was used to train the support vector machine (SVM), float value was set for epsilon for Naive Bayes, number of neighbors of KNN is set as 5, for decision tree the max features set to *in*, the value of penalty of logistic regression is set as elasticnet and for random forest the number if estimates set with. The dataset which contained 373 images with 150 people aged between 60 to 96 was being used. The accuracy for each algorithm were: 78.56% for SVM, 78.12% for logistic regression, 76.58% for KNN, 73.28% for Naïve Bayes, 75.59% for decision tree, and 75.89% for random forest.

B. Challenges in AI-based smart prediction of clinical disease

According to journal Jackins et al. (2020), data mining technique will be the main challenge since it needs to deal with the ever-increasing amount of valuable medical data. The other challenge of this system which is the processing time due to the substantial amount of data used for estimating the performance of the trained data and need to continue it with real-time date in future for estimating the effectiveness of the system and for the future enhancements is the accuracy must be tested with different dataset with applying different AI algorithms to check the accuracy estimation.

According to journal Ahmed & Raheem (2022), there is a main challenge in this study, which is the data quality since the models is rely on the data from Twitter's user. This is because Twitter's user might not offer a fully representative sample of the broader population. Additionally, Twitter's user is come from different culture backgrounds, according to this there will have a dynamic nature of language such as featuring slang, abbreviations, and evolving expressions. Not only the issues of dynamic nature of language, but sentiments also which expressed in tweets may exhibit ambiguity or depend on what is the context. This will cause prediction model's results less accuracy. In the other hand, the other challenge show in this study is the predictive model may not universally apply across diverse regions and cultures since the result may not accurate when the user has a different culture and linguistic background.

According to journal Chakraborty et al.(2022) a conducted analysis is discovered there have knowledge gap in the theoretical guidelines and practical recommendations for creating this lifestyle improvement chatbot. Not only this, chatbot also require a robust dataset for effective model training to obtain and curating a comprehensive dataset with diverse user queries and accurate medical information also as a significant challenge. Other than that, the other challenge is user engagement and satisfaction to the chatbot. In this case, with ensuring the chatbot delivers the accurate result but user-friendly and engaging responses is challenging. Additionally, addressing the proficient in training model, layers and refining the chatbot's conversational abilities in how to describe the predicted result to user are ongoing task for enhanced user satisfaction.

According to journal Kian Siang et al. (2021), there have several challenges are identified in this system. The main challenge in this system is difficulty in data acquisition. This is because the data is overlapping symptoms with COVID-19, this is because clinical signs and laboratory results between dengue fever and COVID-19 is similar. In this case, researcher need to put more effort to find out the significant variable which can distinguishing between this two. Additionally, comparison among various machine learning algorithms on dengue outbreak prediction research is not lack in Malaysia. This lack of research underscores the need for a focused study in this context. Other than that, meteorological, entomological, and socioeconomic factors are crucial predictors for dengue outbreak prediction. In this case, integrating and effectively utilizing these predictors in the system is also a challenge that need to be overcome to have a higher accuracy result.

C. How machine learning contributes to accurate medical diagnosis and predictions

Artificial intelligence methods have brought significant improvements towards the accuracy of disease diagnosing in the medical field. Researchers have constantly presented that AI techniques could produce a better accuracy in early disease diagnosis as compared to humans. Diseases such as brain tumor, Alzheimer's and Parkinson's disease that are difficult to detect by humans (Kahn P, Kader MF, Islam

SMR, Rahman AB, Kamal MS, Toha MU, et. al., 2021) are made possible by AI and the DL-based Convolutional Neural Networking method is the most popular in disease detection. However, recent studies also shows that there are pre-trained models such as EfficientNetB0 model has outperformed other models in accuracy to detect Alzheimer's disease (Savas S, 2022), and combination of different AI algorithms has also revealed a higher accuracy in detecting Parkinson's disease (Lamba R, Gulati T, Jain A., 2022). Although CNN is widely used due to it being fast and accurate in image recognition, research to explore various AI techniques or combining AI techniques is crucial to produce a higher accuracy of disease detection. Therefore, it is important to study different AI techniques that could potentially revolutionize the field of disease diagnosis and predictions.

Artificial Intelligence innovations in the medical field such as AI-related image analysis, lesion determination, health care management have been increasing, and its emergence can be seen in clinical medicine as it is widely implied in the field. The wide application of AI in clinical medicine such as imaging, visualization, segmentations, and extractions is due to its accuracy in disease predictions. Researchers have shown that although it cannot fully replace human providers, it can be supplemented to human providers due to its consistency, speed and repeatability. It is proven AI and clinicians would create a synergistic effect to produce better diagnostic results (Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, et al., 2017) and the accuracy of AI diagnosis has reached a level that matches experts in the field. However, accuracy of AI diagnosis will also experience bottlenecks due to massive amounts of image data in the deep learning neural network can be overwhelming. Thus, it is important that clinical expert/doctor is able combine useful image information with their clinical data to produce a more accurate medical judgement.

Artificial Intelligence in disease detection has been a growing intersection due to its ability to analyze large amounts of medical data and pattern identification features. AI can supplement healthcare professionals in identifying complex and rare diseases to produce better diagnosis accuracy for a better prediction and pattern identification of diseases that might not be apparent to human experts (Nduka et. al., 2019). Accurate diagnosis is a critical task in healthcare to provide patients a precise treatment and outcome. Through the analysis of patient's medical history, genetic data, and other relevant information using AI algorithms, patients can receive the most efficient treatment solution that has better treatment outcomes with lower healthcare costs (Palanisamy et. al., 2019).

There are multiple AI techniques for medical diagnostics such as Random Forest, Artificial Neural Networks (ANN), Naïve Bayes etc. and there are multiple case studies regarding AI-based disease identification to detect diseases such as skin cancer, COVID-19, Alzheimer's disease etc. which are done by researchers from several different institutions, and it has proven that AI is able to generate accuracies of 80% and above in disease detection. However, AI can be limited by data quantity and quality, data and algorithm biases, ethical issues, lack of interpretability and transparency, and lack of generalizability, which would bring disadvantage towards diagnostic results. Therefore, healthcare professionals are advised to invest in research of the technology and approach the technology with caution so that better accuracy, safety and good ethics can be assured.

The presence of Artificial Intelligence has beginning to show impact for clinicians, mainly speed and accuracy increment of image reading; for health systems, in the medical error reduction aspect; for patients, allowing the process of their own data to foster their health. Researchers have found that deep neural network is able to produce high levels of accuracy in radiology for clinicians in pathology, gastroenterology, ophthalmology etc., as for health systems, AI is able to do predictions on key medical outcome though machine vision, wearable sensors, and imaging. Usage of devices such as smart watches and smart phones can assist patients in monitoring their own health.

AI can be integrated in the devices along with sensors or examinations to monitor body conditions. However, despite the promises of AI in medical predictions and diagnosis, there are also several limiting factors of AI, which includes "black box", which refers to flawed algorithms, which has the risk of bringing harm towards patients. Besides, inequities such as human bias which will cause premature mortality. On top of that, data privacy is also an issue due to it being vulnerable to hacking and breached, leading to patient's data getting leaked. Therefore, it is crucial to all sectors and parties should co-operate to produce more AI-medical related papers, ensure better certainty, and reduce AI flaws in diagnosis and predictions.

III. MATERIALS AND METHODS

A. Introduction

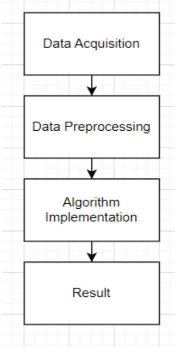

Figure 1: Diabetes prediction model flowchart

Figure 1 shows the flowchart for the diabetes prediction model. It starts with data acquisition. After data acquisition is finished, we will proceed to data preprocessing and a dataset will be generated. The Logistic Regression algorithm will be implemented with Python code and applying the processed data set to the model. A result will be generated after the model optimization, feature selection, and model evaluation.

B. Data Acquisition

- Introduction

A set of suitable data for diabetes predictions regarding several hundred rural African American patients were acquired from Data World (Hoyt, 2018). This dataset is acquired from a study that investigates the common presence of obesity, diabetes and cardiovascular risks, and based on research by Dr. John Wong, who explained that Diabetes Mellitus Type II is strongly associated with obesity. Out of the 1046 African Americans from the rural part of central Virginia, America who are interviewed, 403 subjects were selected for diabetes screening. A positive diabetes diagnosis is identified by numbers above 7 of glycosylated hemoglobin levels (Harrell, 2002).

- Objective

Statistical measurement checking is carried out and it is necessary to enable a better understanding of the data patterns and identify which data attribute is to be cleaned or modified. The data preparation is needed to encompass all the procedures that are required to transform raw data into the final dataset. The primary aim is to produce data that is able to affiliate seamlessly with the requirements of the projects, which will guarantee the efficiency of suitable data handling, cleaning processes, and extraction of relevant attributes from the dataset.

- Description of data

The "diabetes.csv" dataset regarding diabetes detection contains body measurements, lipid levels, heart rate, and personal information of male and female subjects. As mentioned, 403 subjects' data are collected, so there will be a total of 403 records and 19 attributes, which includes, id, chol, stab.glu, hdl, ratio, glyhb, location, age, gender, height, weight, frame, bp.1s, bp.1d, bp.2s, bp.2d, waist, hip and time.ppn. The most important attribute would be glyhb as it guarantees that the subjects' is diabetic. The id is used to uniquely identify the subject's records, chol, stab.glu, hdl are the cholesterol, glucose, and good cholesterol levels respectively in the subject's blood. The ratio attribute defines the ratio of cholesterol to good cholesterol. Bp.1s, bp.1d, bp.2s and bp.2d are the systolic and diastolic blood pressure measurement of the subjects. Height and weight can be used to calculate the BMI, as well as the hip and waist can be used for calculating hip and waist ratio that could become the predictors of diabetes. Location, age, and gender are the personal data of subjects.

C. *Data preprocessing*

The data preprocessing was done with several simple operations. Any subject records without glycolates hemoglobin, glyhb were all excluded as it can be used to define whether the subject is diabetic or not. If the glyhb level was 7 or greater, they will be labelled as "Diabetes = yes", other than that will be "Diabetes = No diabetes". Then, the outcomes of "Yes" and "No diabetes" will be converted into binary numbers 1 and 0 respectively, total of 390 records have been produced. All 390 subjects will be grouped by age from the youngest to oldest and the id of subjects will be converted into numbers 1 to 390, labelling as patient_number for unique identifications of the records. The chol and hdl is labelled as Cholesterol and HDL Chol, initial "ratio" attribute is labelled as Chol/HDL ratio for Cholesterol to HDL Chol ratio calculations. A new BMI attribute is added for BMI calculations of subject's weight (lb) and height (in) using the formula of "weight(lb) / [height(in)]2 * 703". The hip to waist ratio attribute, labelled as Waist/hip ratio is also added. The bp.1s, bp.2s attributes are classified and labelled as Systolic BP, as well as bp.1d, bp.2d, labelled as Diastolic BP. "Female" and "male" values of the gender attribute are converted into binary values, 0 and 1 respectively. Unnecessary attributes such as Location are removed as it will not affect the results of the diabetes forecasting.

D. *Algorithm Implementation*

- Logistics Regression

Logistics Regression (LR) is an ML approach and supervised machine learning models that are used for prediction and classification problems. The class labels in the prediction and classification models are needed for a given set of input variables. LR can estimate the probability of the outcome or something happening based on the given dataset of independent variables. The outcome of LR is a probability so the dependent variable is bounded between 0 and 1 (true and false) values (IBM, n.d.). There are some examples of scenarios of how LR is applied in different activities such as identifying the data anomalies for predicting fraud, illness, and disease prediction, predicting the churn in different functions of the organization by analyzing the specific behaviors, and the Binary LR approach used for predicting the email spam or not spam. In this research, the LR algorithm was trained and tested with diabetes data sets for predicting diabetes.

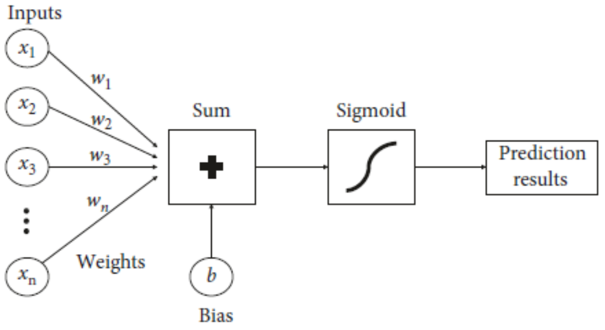

Figure 2: Flowchart of architecture of logistics regression (Antor et al., 2021)

Figure 2 shows the flow of function of the logistic regression model. The independent and dependent variables are determined in this model. It predicts probability and establishes decision boundaries using the sigmoid function. There are different parameters being used such as l2 penalty or l1 penalty, different C parameter values for the fine-tuning (Antor et al., 2021).

- Purpose

The outcome of predicting diabetes is about 0 or 1 (true or false) to represent the absence or presence of diabetes. Based on the outcome that wants to be achieved, the Logistic Regression (LR) is a well-suited algorithm for binary classification tasks which are 0 and 1. It is more appropriate for predicting the likelihood of diabetes and is easy to get and understand the answers.

Next, the interpretability of the model is important, especially in medical applications to easily understand the factors that contribute to the predictions. LR provides the benefit of easily understanding and explaining how each variable affects the probability of diabetes. It can save time and resources and put more effort into other aspects.

- Parameters and variables

There is one parameter being changed and experimented on to see the result of accuracy. The penalty value of the model will be set with ‘l1’ and ‘l2’ respectively to experiment on the diabetes data set. The penalty is used to handle and prevent overfitting and minimize the model’s generalization error. This approach can avoid the more complex model learning process to prevent overfitting (Melanee Group, 2023). While the remaining parameters will remain the default values such as solver with the value 'liblinear', value of C is 1, value of tol is ‘1e-4’, value of fit_intercept is ‘True’ and value of intercept_scaling is 1.

The main two independent variables which are ‘Age’, ‘Glucose’ with the dependent variable ‘Diabetes’ are selected to understand how these two features relate to the ‘Diabetes’ variable. Other than that, all the variables will be put into the model to evaluate the correlation relationship with the ‘Diabetes’ variable.

IV. Result and Discussion

A. Result

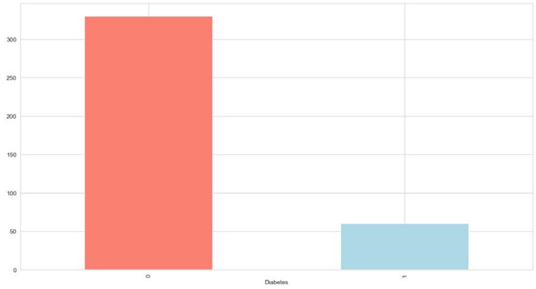

Figure 3: Number of Diabetic Patients and Non-Diabetic Patients

Figure 3 shows a bar chart of numbers of diabetic patients and non-diabetic patients. The x-axis represents the diabetic outcomes of patients, which are identified with 0, no-diabetes, and 1, diabetes. The y-axis represents the number of patients for each outcome. The number of patients, represented by the red colored bar, which are non-diabetic (0), has a total number of 330 patients. The number of patients, represented by the blue colored bar, which are diabetic (1), has a total number of 60 patients.

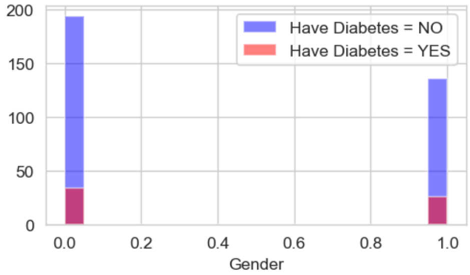

Figure 4: Number of patients that are diabetic or non-diabetic for each gender

Figure 4 shows a histogram of the number of patients for each gender, and whether they are diabetic or not. The x-axis of the histogram represents the genders of the patient, where female gender is identified with number 0.0, and males are identified with number 1.0. The y-axis will represent the number of patients for each gender. Blue colored bar indicates the number of patients that are non-diabetic, and red colored bar indicates the number of patients that are diabetic. The number of diabetic female patients are 34 and 194 non-diabetic patients, adding up to a total of 228 female patients from the dataset. On the other hand, the number of non-diabetic male patients is 136, whereas 26 of male patients are diabetic, adding up to a total of 162 male patient records from the dataset. The results prove that gender of patients is not the main factor of diabetes.

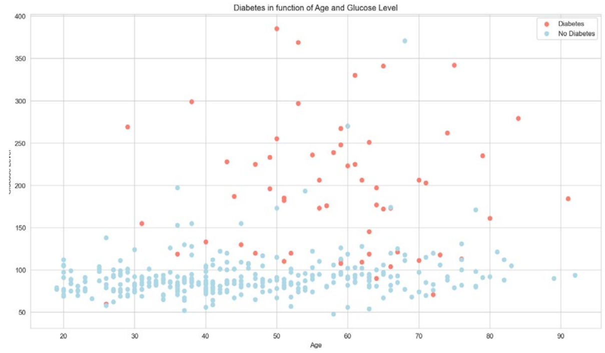

Figure 5: Functions of Age and Glucose Levels that causes Diabetes

Figure 5 shows a scatter plot between the functions of age and glucose levels that causes diabetes. X-axis of the graph represents the age of patient and Y-axis of the graph represents the glucose levels in the patients’ blood. The red colored dot in the graph will be patients that have diabetes and blue colored dots are the patients that do not have diabetes. The outcome of the graph indicates that the majority of the non-diabetic patients have glucose levels lower than 100 and aged below 70. Most of the diabetic patients have glucose levels that are between 100 to 300 and aged between 40 to 70. Patients that aged older (>50) are more prone towards diabetes, whereas patients that are aged younger (<40) are less prone towards diabetes. As for glucose levels, patients that have glucose levels above 100 are more likely to have diabetes, whereas patients that have glucose levels below 100 are less likely to have diabetes. The result of this graph explains the increase in age, along with higher glucose levels are positively correlated with an elevated risk of diabetes.

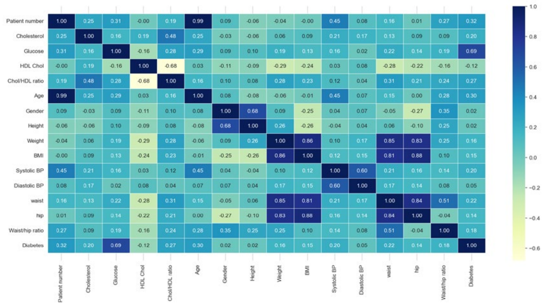

Figure 6: Correlation Matix of the dataset

Figure 4 shows the correlation matrix between each pair of attributes of the dataset. A strong linear relationship is indicated by coefficients that are close to 1 or -1, whereas a weak linear relationship is indicated by coefficients that are close to 0. A positive coefficient, which is closer to 1, indicates a positive correlation, on the other hand, a negative coefficient, which is closer to -1 indicates a negative correlation. The highest positive correlation, 0.99 correlation coefficient, is between patient_number and age, as the patient_number from 1-390 are arranged ascendingly according to age. The highest negative correlation, -0.68 correlation coefficient, between Chol/HDL ratio and HDL chol, which explains when Chol/HDL ratio increases, the HDL chol will decrease. The weakest correlations are coefficients between BMI and patient_number as well as coefficients between hip and age, which have 0 correlation coefficient, explains that the coefficients have no relationships at all.

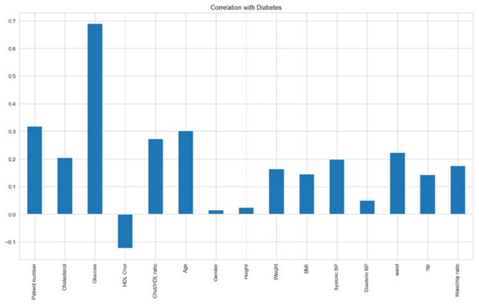

Figure 7: Correlation with Diabetics

Figure 7 shows a graph that indicates the correlation with diabetes. Similarly, the theory of this graph will be the same as Figure 4, except it will be targeted towards the diabetes outcome of patients. The highest correlation coefficient is 0.68, whereas the lowest correlation coefficient is -0.12. The highest positive correlation is between glucose and diabetes, which indicates that high glucose levels in patient's blood are the main risk factor of diabetes. The highest negative correlation with diabetes is HDL chol, but due to it being very near to 0, that will explain that lower HDL chol will only have a slight effect on diabetic outcomes of patients. The coefficient that is the weakest with diabetes is gender and height, with a correlation coefficient of 0.2, which indicates that height and gender does not have any relationship with the diabetic outcomes of the patients.

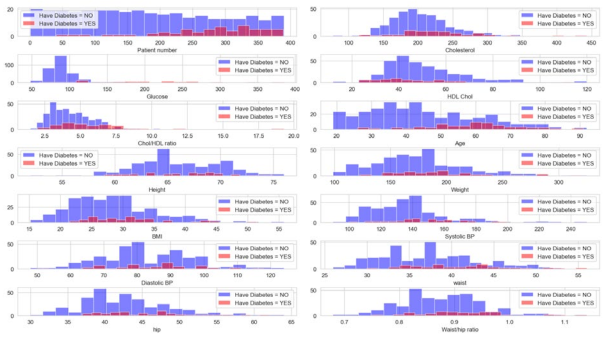

Figure 8: Histograms of each variable in the data set

Figure 8 shows the histograms for each variable in the dataset, which could possibly define the main risk factors of diabetes. The y-axis will represent the number of patients for all histograms, and the x-axis of the histograms represents the variables available in the dataset, excluding diabetes. Blue colored bar shows patients that are non-diabetic, and red colored bar shows patients that are diabetic. The one of the more significant findings from Figure 6 is age, where patients surpass the age of 50 are more prone towards diabetes. The histogram explains that when patients get progressively older, the higher the risk of getting diabetes when they are not aware of their health. Another significant finding from Figure 6 will be glucose levels of patients, where patients with glucose levels higher that 100 has higher risk of getting diabetes. This indicates that the patients' blood glucose levels have a positive correlation with risk of getting diabetes. There is also an inverse relationship between diabetes and HDL cholesterol which is proven in Figure 6. The HDL chol histogram shows that patients are more prone towards diabetes when they have lesser levels of HDL cholesterol. The ratios of hip/waist and chol/hdl is also another significant find that shows when the ratios progressively increase, the number of diabetic patients also increase. As for the other variables in the dataset, they have slight to no effect towards diabetic outcomes of patients. due to it having inconsistent number of diabetic patients across the x-axis of the histogram.

Penalty	Training Accuracy	Testing Accuracy
l1	93.04%	91.45%
l2	93.46%	90.60%

Table 1: Results of different penalty

TABLE I shows the result of using different penalty values. There are two different penalties used which are l1 and l2. Based on the table, we found that the training accuracy of l2 is more than the accuracy of l1 but the accuracy of testing of l1 is more than l2. This result shows that the l2 penalty might be better in fitting the training data while the l1 penalty is performing better on unseen data and it is an important factor in model evaluation. In our perspective, the l1 penalty is more suitable as it is better with a high testing accuracy. Overall, the accuracy of logistic regression model for predicting diabetes is more than 90% which means that both results have achieved good performance.

B. Discussion

In various healthcare sectors, the utilization of machine learning among the sectors has seen a rapid growth as Machine Learning is able to generate good performances, fairly accurate prediction outcomes, and accelerated results. Diabetes, a type of chronic disease, often due to high blood glucose levels that due to insufficient insulin hormones to regulate glucose in the blood (WHO, 2023), and it has been a plaguing effect towards US citizens in recent years. In this study, application of machine learning is crucial to tackle diabetes in-terms of predictions, diagnosis, and cure.

Logistic Regression, a supervised learning algorithm of machine learning, was employed as a predictive modeling technique to determine the risk of getting diabetes according to several contributing factors. The dataset that includes, the body measurements, blood measurements and biomarkers, and demographic information from a sample of rural African Americans from central Virginia. Training of the model took place using a subset of the data, and the performance of the model is evaluated through an independent test set. The results then found out that variables that are highly correlated to diabetes includes age, HDL cholesterol and glucose levels in blood. The logistic regression model is able to produce a high predictive accuracy, and it indicates that the model can differentiate between patients that are diabetic, and non-diabetic.

These results have proven that logistic regression can be used as a suitable predictive tool for diabetes according to readily available and clinical data. However, it is also important to acknowledge the limitations, such as acquired dataset that is too small, which could possibly cause problems like limited statistical power, sampling bias, and unreliable results. Future research can be carried out to emphasize on integration of additional variables and consider longitudinal data to enhance the robustness and generalizability of predictive models for diabetes.

V. Conclusion

A diabetes prediction model is more crucial than ever to be an assistance to people to help them understand their body health more and can prevent diabetes. In this case, the study shows that people who surpass 50 are not aware about their health and this group of people will have a high risk of getting diabetes. This model used logistics regression algorithm to train the dataset to get the result of the people who have diabetes or don't have diabetes. In this case, this model has performed well with the accuracy more than 90%. However, furthermore research and version of dataset is the key to enhance the accuracy of the model.

Acknowledgment

We would like to thank Mr Robert for sharing his diabetes dataset available publicly on Data World and we would also thank Mr Srikanth for sharing his logistics regression model available publicly on GitHub for our reference. Finally, thanks to Dr. Adeline as our mentor and always providing suggestions for us to conduct good research.

References

[1] Institute for Health Metrics and Evaluation. (n.d.). *Global diabetes cases to soar from 529 million to 1.3 billion by 2050.* https://www.healthdata.org/news-events/newsroom/news-releases/global-diabetes-cases-soar-529-million-13-billion-2050

[2] Khalid, A., Mehmood, A., Alabrah, A., Alkhamees, B. F., Amin, F., AlSalman, H., & Choi, G. S. (2023). Breast cancer detection and prevention using machine learning. *Diagnostics*, *13*(19), 3113. https://doi.org/10.3390/diagnostics13193113

[3] Rabiei, R. (2022). Prediction of Breast Cancer using Machine Learning Approaches. *Journal of Biomedical Physics & Engineering*, *12*(3). https://doi.org/10.31661/jbpe.v0i0.2109-1403

[4] Purushottam, Saxena, K., & Sharma, R. (2016). Efficient heart disease prediction system. *Procedia Computer Science*, *85*, 962–969. https://doi.org/10.1016/j.procs.2016.05.288

[5] Bhatt, C., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective heart disease prediction using machine learning techniques. *Algorithms*, *16*(2), 88. https://doi.org/10.3390/a16020088

[6] Neelaveni, J., & Devasana, G. (2020). Alzheimer Disease Prediction using Machine Learning Algorithms. *IEEE*. https://doi.org/10.1109/icaccs48705.2020.9074248

[7] Zhang, Z., Chuah, J. H., Lai, K. W., Chow, C., Gochoo, M., Dhanalakshmi, S., Wang, N., Bao, W., & Wu, X. (2023). Conventional machine learning and deep learning in Alzheimer's disease diagnosis using neuroimaging: A review. *Frontiers in Computational Neuroscience*, *17*. https://doi.org/10.3389/fncom.2023.1038636

[8] Dashtipour, K., Taylor, W., Ansari, S., Zahid, A., Gogate, M., Ahmad, J., Assaleh, K., Arshad, K., Imran, M. A., & Abbai, Q. (2022). Detecting Alzheimer's disease using machine learning methods. In *Springer eBooks* (pp. 89–100). https://doi.org/10.1007/978-3-030-95593-9_8

[9] IBM. (n.d.). *What is Logistic regression? | IBM*. https://www.ibm.com/topics/logistic-regression

[10] Antor, M. B., Jamil, A., Mamtaz, M., Khan, M. M., Aljahdali, S., Kaur, M., Singh, P., & Masud, M. (2021). A comparative analysis of machine learning algorithms to predict Alzheimer's disease. *Journal of Healthcare Engineering*, *2021*, 1–12. https://doi.org/10.1155/2021/9917919

[11] Melanee Group. (2023, May 21). A comprehensive analysis of hyperparameter optimization in logistic regression models. *Medium*. https://levelup.gitconnected.com/a-comprehensive-analysis-of-hyperparameter-optimization-in-logistic-regression-models-521564c1bfc0

[12] World Health Organization. (2023, April 5). *Diabetes*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/diabetes#:~:text=Diabetes%20is%20a%20chronic%20disease,hormone%20that%20regulates%20blood%20glucose.

[13] Harrell Jr, F. E. (2002, December 27). Diabetes Dataset. https://hbiostat.org/data/repo/diabetes

[14] Hoyt, R. (2021, June 6). *Diabetes Prediction - dataset by informatics-edu*. data.world. https://data.world/informatics-edu/diabetes-prediction/workspace/file?filename=Diabetes_Classification.xlsx

[15] Ghaffar Nia, N., Kaplanoglu, E. & Nasab, A. (2023). Evaluation of artificial intelligence techniques in disease diagnosis and prediction. *Discover Artificial Intelligence 3*(1):5. doi: 10.1007/s44163-023-00049-5.

[16] Topol, E. J. (2019). High-performance Medicine: the convergence of human and artificial intelligence. *Nature Medicine, 25*(1):44-56. doi: 10.1038/s41591-018-0300-7

[17] Ramudu, Kama & Mohan, V & Jyothirmai, D & Prasad, D & Agrawal, Ruchi & Boopathi, Samapth. (2023). Machine Learning and Artificial Intelligence in Disease Prediction: Applications, Challenges, Limitations, Case Studies, and Future Directions. *Contemporary Applications of Data Fusion for Advance Healthcare Informatics*, 297-318. doi: 10.4018/978-1-6684-8913-0.ch013.

[18] Liu, C., Jiao, D., Liu, Z. (2020). Artificial Intelligence (AI)-aided Disease Prediction. *Bio Integration, 1*(3):130-136. doi: 10.15212/bioi-2020-0017

[19] Ahmed, N., & Raheem, M. (2022). Sentiment Prediction on COVID-19 Vaccination Reviews. Journal of Applied Technology and Innovation, 6(3), e-ISSN: 2600-7304.

[20] Kian Siang, T., Ramachandran, C. R., & Meskaran, Dr. F. (2021). Dengue disease prediction using machine learning algorithms: a review. Journal of Applied Technology and Innovation, 5(4), e-ISSN: 2600-7304.

[21] Gandhi, G. M., Singh, V., & Kumar, V. (2019). IntelliDoctor - AI based Medical Assistant. IntelliDoctor - AI Based Medical Assistant. https://doi.org/10.1109/iconstem.2019.8918778

[22] Jackins, V., Vimal, S., Kaliappan, M., & Lee, M. Y. (2020). AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *The Journal of Supercomputing*, *77*(5), 5198–5219. https://doi.org/10.1007/s11227-020-03481-x