International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Exploring Analytical Insights for Understanding and Managing Type2 Diabetes

Leelambika KV[1], Dr. Shamugarathinam G[2]

Submitted: 26/01/2024 **Revised**: 04/03/2024 **Accepted**: 12/03/2024

Abstract: These The frequency of diabetes is surging at an astonishing and concerning pace, and in a survey statistics, by 2040, 640 million people will be a diabetic worldwide, out of which 90% having Type2 diabetes and 10% with Type1 diabetes and Gestational diabetes. World Health Organization (WHO) reported that India ranked second position in the diabetes prevalence. This paper focus on the systematical review and critical analysis of various applications and implications of predictive analytic in the context of Type 2 Diabetes management, highlighting its role in shaping proactive healthcare approaches and the life quality of the affected individuals by this condition get improved. As the prolonged state, diabetes heightens the susceptibility of patients to renal complications, coronary heart diseases, and vascular disorders. Therefore, through a comprehensive examination of studies, methodologies, and real-world implementations, this paper underscores the transformative impact of harnessing predictive analytics for the holistic understanding and management of Type 2 Diabetes.

Keywords: Prediction, Data Analytics, Machine Learning, Decision making, Classifiers, Accuracy, Type 2 diabetes

1. INTRODUCTION

A. Overview of Diabetes

Diabetes is a multifaceted and intricate metabolic disorder, arising from inadequate insulin secretion by the pancreas' beta cells. The primary factor contributing to this condition is that irregular food intake irrespective of time, smoking and life style habits. There are four types of diabetes: Type1, Type2, gestational diabetes and pre-diabetes. The immune system responsible for protecting the body from harmful infections, mistakenly attacks and destroys the beta cells in the pancreas. So, the pancreatic beta cells produce little or no insulin. This causes the flow of glucose level is high in the blood stream and this state is refereed as Type1 diabetes. While treating this, patients are given with insulin injections. Hence, it is called as Insulin Dependent Diabetes Mellitus. It can manifest at any age, often seen in children, adolescents, or young adults and requires lifelong insulin therapy.

Type2 diabetes stands as the most prevalent form of diabetes in which blood glucose level is too high. The pancreatic beta cells produces insulin, but the body cells are no longer using the hormone efficiently. This leads to the higher flow of glucose in the blood stream, causes this T2D. And this is common in older adults or 45+ ages. The cause of type2 diabetes is linked often with lifestyle factors, genetics and obesity. T2D can be managed by making life style changes, taking medications, exercises and regular check-ups.

Gestational diabetes is common in the pregnant women and causes high blood sugar which affects the pregnancy as well as the baby's health. Pre-diabetes state is in which, the level of the sugar in the blood stream is in abnormal range, but couldn't categories as diabetic state. If untreated, it leads to serious health issues and causes type2 diabetes. By following the healthy life style and exercise, pre-diabetes can be curable. This paper focuses on the predictive models and the clinical factors involving in it. The parameters to diagnosis the diabetes are demographics (age, gender, and total quantity) and biological markers (HbA1c, BMI, and BP).

B. Importance of Analysis in healthcare

This research paper involves the data analytics in healthcare domain. The process of analyzing the raw data, and find the knowledge and patterns to draw conclusions by identifying the insights of the data is known as Data analytics. It plays a significant role in the healthcare domain in order to improve the patient care by more accurately diagnosing the disease by examining the data sets, proposing a decision for the treatment and preventative measures. The health data to be collected by electronic health records (EHR) or by e-prescription services or by patient portals or by health related software apps. There are different types of health care analytics: Descriptive analytics is used for analyzing the past data to draw comparisons and discover patterns. Predictive analytics uses both historical and current data to gain insights into the data. Prescriptive analytics make predictions about future outcomes. Machine learning in data analytics constitutes model-building automation for data analysis.

Metrics used for Diabetes diagnosis

In predicting or diagnosing diabetes, various metrics or indicators can be used to assess an individual's risk or condition.

These metrics are often derived from medical tests, measurements, and data analysis. Here are some common metrics used to find and assess diabetes in the table 1:

Some of the other characteristics are family history, age and gender, physical activity level, genetic factors and the history of health. This paper is further organized in such a way that section II

[1] *School of CSE, Presidency University, Bangalore – 560064, INDIA*
ORCID ID : 0009-0009-6379-1886
[2] *School of CSE, Presidency University, Bangalore – 560064, INDIA*
ORCID ID : 0000-0002-8630-7470
** Corresponding Author Email: vleelambika@gmail.com*

as related work, section III as data set collection, section IV as steps in predictive model and section V as conclusions.

2. RELATED WORK

According to the literature, researchers have employed machine learning algorithms to analyze Type 2 Diabetes (T2D) disease. Haixia Shang et al (2021)[1], identified the disease genes causing T2D. Weighted page rank algorithm is used to find the diabetes genes, which assigns higher rank to the set of genes and comparatively lower rank to other set of genes. They constructed bilayer network for proteomics and transcriptomics, respectively defined as protein to protein relation and gene to gene relation by DMI (Differential Mutual Information). In addition, gene to protein relation of DMI was found to examine the regulatory relationship. The limitation of this gene regulatory relationship is that, it gets affected by environmental factors such as diet, humidity, temperature, oxygen levels, cycles of light and the mutagens presence.

Liying Zhang et al (2021)[2], aim is to enhance the ability to recognize T2D effectively by Joint Bagging-Boosting Model. Random Forest in bagging model, classifies the data with row sampling and feature sampling with replacement, using multiple decision trees (DT). Gradient and Extreme Gradient Boosting algorithms are used. Gradient Boosting is used to find the best next model from the previous and minimizes the overall prediction error. In order to find the precise model, weight of the variables are considered in the Extreme Gradient Boosting and gets processed in the decision tree. Some of the demographic parameters involved are gender, age, marital status, Education level and the anthropometric information considered are waist to hip ratio, heart rate, and blood pressure.The limitation of RF is that, it is fast to train the data but low to create predictions, once they are trained. Gradient Boosting is more accurate than RF, but sensitive to outliers. XGBoost does not perform well on sparse and unstructured data.

Table 1. Key metrics for diabetes detection

Key Performance Indicating metrics	***Non Diabetic Range***	***Diabetic Range***	***Key Performance Indicating metrics***	***Non Diabetic Range***	***Diabetic Range***
Fasting Blood Glucose Test	70 - 99 mg/dL	100 mg/dL	Body Mass Index (BMI)	18.5 - 24.9 kg/m²	Overweight - 25 kg/m² or higher Obese - 30 kg/m² or higher
Oral Glucose Tolerance Test	Less than 140 mg/dL	200 mg/dL or higher	Waist Circumference	Men: <120cm (40 inches) Women: <88cm (35 inches)	Men: ≥ 102 cm (40 inches) Women: ≥ 88 cm (35 inches)
Hemoglobin A1c (HbA1c) Test	< 5.7%	6.5% or higher	Blood Pressure	< 120/80 mmHg	≥ 130/80 mmHg (Elevated) or ≥ 140/90 mmHg (Hypertension)
Fasting Insulin Levels	Normal range varies	Elevated levels may indicate insulin resistance	Cholesterol Levels	HDL Cholesterol: Men: ≥ 40 mg/dL Women: ≥ 50 mg/dL	HDL Cholesterol: Men: < 40 mg/dL Women: < 50 mg/dL Triglycerides: ≥ 150 mg/dL

Triglycerides: < 150 mg/dL	Total Cholesterol: ≥ 200 mg/dL
Total Cholesterol: < 200 mg/dL	

Nada Y. Philip et al(2021) [3], identified a associations between patients biological markers and complications by proposing a suite for T2D in data analytics. There are three stages: classification, risk prediction and response for the treatment. In the profile classification of the patient, investigations are made based on the association of different patients. Cox's Proportional Hazards model (CPH) is used to estimate risk factors and 10-Fold Cross Validation method is used for the treatment of response prediction. The attributes taken for consideration are demographics (age, gender, and total quantity) and biological markers (HbA1c, BMI, and BP). The violation of the proportional hazard assumption in the Cox Regression model can indeed lead to the creation of a false model that may not accurately represent the relationship between the predictor variables and survival time. The 10 Fold CV model needs to be trained K times at the validation step and it requires higher computational costs.

Benjamin Lobo et al (2021) [4], developed a drive of Continuous glucose monitoring CGM for every 2hours, in which RMSE (Root Mean Squared Error) was calculated. All-motifs algorithm and Classify algorithm are used to find the set of Daily profile representations. A set of 8 Candidate daily profiles and motifs (finite set of profiles) are identified by the All-motifs and Classify algorithms respectively. These motifs act like decorative images or designs that capture patterns from recurring forms. The features of the data sources are BMI, CLC, MDI, NR and SAP. Rather than real-time blood glucose levels, the measurement of interstitial glucose levels are found as a limitation.

Sumeet Kalia et al (2022) [5], investigates and estimates the diabetes by formulating the marginal structural model, by combining the drug therapies such as metformin, sulfonylurea and Sodium-Glucose Co-transporter (SGLT-2i). Metformin improves the body to take insulin and lowers the blood sugar level. Sulfonylurea is a drug used to increase the secretion of insulin and lowers the sugar level in the bloodstream. SGLT2i is an inhibitor which protects the heart and kidney failures by reducing the blood glucose. They considered two sets of cohorts such as naive and drop-in cohorts. The naive treatment means the patients are not yet taken treatment for a disease. And treatment drop-in are the patients started the medication but waiting for the outcomes. Machine learning pipelines are used for the 3 sections: longitudinal cohort, covariate balance, hypothetical prediction. The clinical parameters considered are BP, Weight, Hemoglobin A1c, lipid, ACR. This is considered as blackbox due to complexity.

Harleen Karur et al (2023) [6], the goal is to develop predictive models that can classify patients as diabetic or non-diabetic based on various risk factors, contributing to the global concern of rising diabetes cases. The specific models utilized are: SVM-Linear (Linear Kernel Support Vector Machine), Radial Basis Function K-SVM (Radial Basis Function Kernel Support Vector Machine), k-NN (K-Nearest Neighbors), ANN (Artificial Neural Network), MDR (Multifactor Dimensionality Reduction). By using R data manipulation tools, the above models are implemented. The Pima Indian diabetes datasets are anayzed in order to identify trends, patterns, and risk factors associated with diabetes. The ultimate aim is to enhance the accuracy of diabetes prediction through the application of these machine learning models. The potential limitations could include the dataset's representative and diversity, as well as the risk of over-fitting with complex machine learning models.

3. Choosing A Machine Learning Algorithm For A Use Case

For the supervised learning, with the data sets, either by using regression or classification, the use cases is solved. The classification problems are resolved using

- Logistic regression (LR),
- Decision tree (DT),
- Random Forest (RF),
- XGBoost
- K-NN
- Weighted SVM
- Feed Forward Neural Networks (FNN)

How to decide the particular algorithm is used? Each and every algorithm has its own advantages and disadvantages. What is the strategy used to find the best model? The training of large data sets takes much time, while deciding the algorithms. The data visualization plays a major role in this concern. The seaborn libraries are used in the above case in pairplot. Initially, the datasets are classified into independent and dependent features. The bivariant data are plotted in the pairplot, in order to choose the type of classification algorithm to be used. While considering the datasets, if there is no overlapping or less and can able to draw a best fit line to divide the particular points clearly, then logistic regression is used. The accuracy is high; so for linear classification, logistic regression is used. In case of non-linear classification, that is, if there is an overlapping in the points, then Decision Tree or Random Forest or KNN is used.

A. Logistic Regression

The sigmoid function is the representation of logistic regression models and is given by and is explained in the figure 1 as follows:

$$g(z) = \frac{1}{1 + e^{-z}}$$

The value of threshold is depending on two factors, such a

precision and recall. Theoretically, the value of both precision and recall are 1. But, in the real case this trade-off is not possible.

B. DECISION TREE

A decision tree is most commonly used to find the classification and regression. It is used for making the decisions. The classification tree is used to classify the target patients are survived or not alive. The regression tree gives the continue monitoring of the target patients.

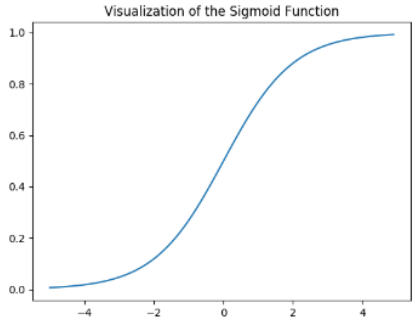

Fig. 1. Sigmoid function of LR

Regression = sum(y — prediction)²

Classification:

$G = SUM\ (p_k * (1 - p_k)$

Here, p_k is same class inputs proportion present in a particular group.

Lets consider, two features, Feature_A and Feature_B and the goal is to predict either class 0 or 1.

IF Feature_A <= Threshold_A:

IF Feature_B <= Threshold_B:

Prediction: Class 0

ELSE:

Prediction: Class 1

ELSE:

Prediction: Class 1

Where, Feature_A and Feature_B are the feature values and

Threshold_A and Threshold_B are the decision thresholds for the features.

C. Random Forest (RF)

Random forest is a technique of ensemble , which involves the combination of number of models as given in fig 2. For making predictions, numbers of models are considered instead of an individual model. There are two types of methods, bagging and boosting.

Bagging: The sample training data with replacement is given, to create different training subsets. The majority voting is stated as the final output. **Boosting:** The sequential model is created to combine the weak learners into strong learners and in such a way that the final model with highest accuracy.

This is given by the formula as follows:

Random Forest = DT (base learner) + bagging (Row sampling with replacement) + feature bagging (column sampling) + aggregation(mean/median, majority vote)

D. XGBOOST

A machine learning algorithm, describes about the distributed gradient boosted decision tree is referred as Extreme Gradient Boosting (XGBoost) as shown in fig 3.

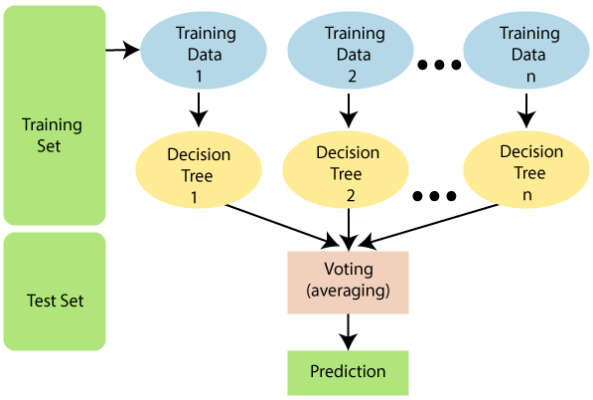

Fig. 2. Work flow of Random Forest

This is given by the formula as follows:

Random Forest = DT (base learner) + bagging (Row sampling with replacement) + feature bagging (column sampling) + aggregation(mean/median, majority vote)

E. XGBOOST

A machine learning algorithm, describes about the distributed gradient boosted decision tree is referred as Extreme Gradient Boosting (XGBoost). XGBoost is used for the problems such as regression, classification and ranking.

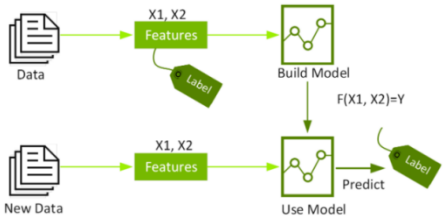

Fig. 3. Workflow of XGBoost

The precision score of the individual decision tree is given by

$$\hat{y}_i = \sum_{k=1}^{K} f_k \epsilon F$$

Where, k - number of trees,

f - Functional space and F - possible set available

F. K-NEARST NEIGHBOR (KNN) CLASSIFIER

The k-Nearest Neighbors (KNN) algorithm classifies new data points based on the similarity measure between these new data points and the earlier data points in the training dataset. KNN is measured by distance function and is given by

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

G. WEIGHTED SUPPORT VECTOR MACHINE(SVM)

SVM gives a better results on smaller datasets and is categorized as supervised algorithm. SVM can be used for regression and classification. Both SVM and logistic regression are used to find the hyperplane, but LR is a probabilistic approach whereas SVM is based on statistical approach. The hypothesis function (h) for SVM is given by

$$h(x_i) = \begin{cases} +1 & if\ w \cdot x + b \geq 0 \\ -1 & if\ w \cdot x + b < 0 \end{cases}$$

This weighted SVM allows differential weight allocation based on the classes which are imbalance and more critical than others. So this mechanism provides different importance to each classes.

H. FEED FORWARD NEURAL NETWORK (FNN)

FNN is used to predict the category or class of the input data. It is a type of Artificial Neural Network, also known as MLP Multilayer Perceptron, belongs to supervised learning algorithms. It comprises of input, hidden and output layers. Each layer consist of neurons (nodes), receives inputs and performs a weighted sum and an activation function, produces an output as predictions. The feed forward model is given by,

y = f* (x)

Where, input x is assigned to category y.

4. METHODOLOGY

In this paper, the different phases get involved in analyzing the diabetes datasets. The framework of machine learning and a deep learning techniques are illustrated in Figure 4. Colab is used for the entire implementation. The packages such as NumPy, Pandas, Scikit, tensorflow, keras and Matplotlib are used to analysis the data.

4.1 Datasets:

The PIMA dataset is used for analysis. It consist of 768 observations / rows and 9 variables. The variables or attributes are pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age and outcome.

Description about the PIMA dataset:

Pregnancies - Number of times a woman has been pregnant; Glucose - By Oral glucose tolerance test, Plasma glucose concentration of 2 hours is taken; Blood Pressure - Diastollic blood pressure in mmHg; Skin Thickness - Triceps skin fold thickness in mm; Insulin - 2hour serum insulin in mu U/ml; BMI - Body Mass Index; Diabetes Pedigree Function - diabetic score based on family history; Age - Age in years; Outcome - 0:Non-diabetic, 1:Diabetic patient.

The PIMA diabetic dataset is a CSV file and read the file using Pandas function.

4.2 Exploratory Data Analysis:

While describing the data, spread across the table, it has been observed that the minimum value of some variable is zero, but cannot be so in medical grounds. So, such values are replaced with median/mean value depending on the distribution in the data cleaning process. In the similar case, maximum value also so high as 846 in the insulin. This can be treated as outliers.

4.3 Data Cleaning:

The critical step in the data analysis is the data cleaning. The identification of the missing values, handling of the duplicated data, dealing with outliers, handling of irrelevant data and the removal of unnecessary variables based on the requirement plays a vital role in the modeling and to achieve best accuracy rate. As the new data is collected then regularly revisit and update the data cleaning is essential.

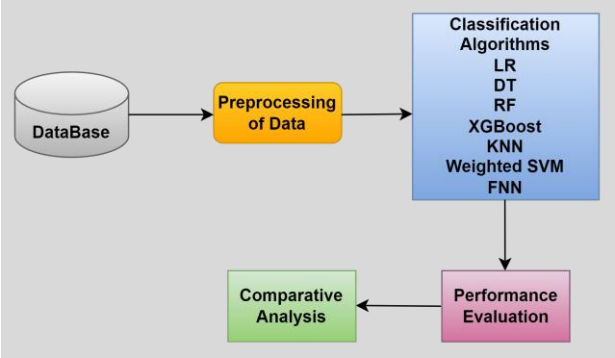

Fig. 4. Framework of the model

4.4 Data Visualization:

The plotting of different graphs, ensures about the data that extracted is suited for the specific model. In order to check where the data set is balanced, then Count Plot is used. The histogram is used to verify if the data is normally distributed or not. The box plot is used to analyse the distribution of data and to look into the outliers; the scattter plot is to understand the relationship between the variables.

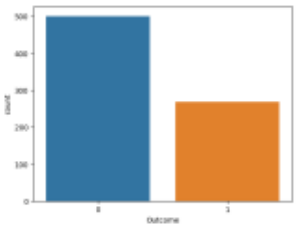

Fig. 5. Outcome Variable - Diabetic vs Non-diabetic measure (Imbalanced data)

4.5 Feature Selection:

Pearson's Coorelation Coefficient:

It is a measure of the linear relationship between the two variables.

It is used to evaluate the strength of association of the two continuous variables. The measure of coefficient ranges from -1 to 1, where, 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship.

In the heatmap figure, it is observed that, Glucose, BMI and age are correlated closely to the Outcome variable. Blood Pressure, Insulin and DiabetesPedigreeFunction are less related to the outcome variable.

4.6 Build the Classification Algorithms:

The PIMA diabetic datasets is analyzed with machine learning and deep learning models. The analysis report of the models are shown in the below tables. By using the PIMA datasets, the models such as LR, DT, RF, XGBoost, KNN, Weighted SVM and FNN are analyzed.

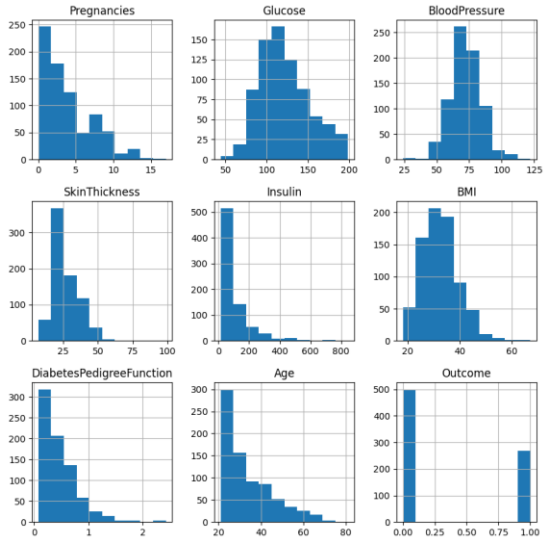

Fig. 6. Histogram of each variable in the datasets

Observation : The variables such as Glucose and Blood Pressure are normally distributed and the remaining are skewed

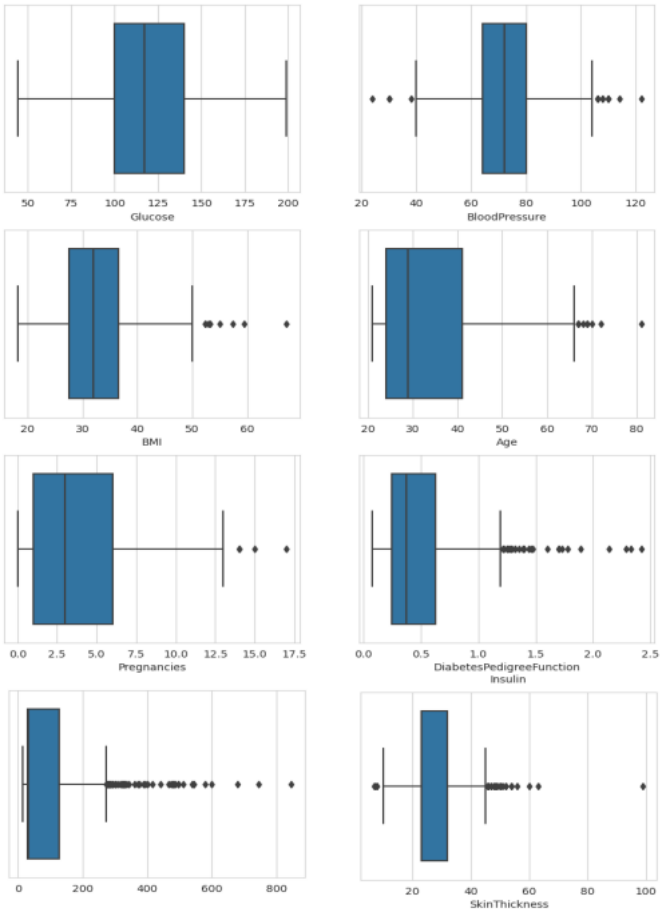

Fig. 7. Box Plot - to find outliers

The most relevant features are selected, the outliers are handled and are divided into two portions for training and testing. The datasets are splitted in such a way that, 80% for training and 20% for testing. Then, the models are build to determine the performance metrics.

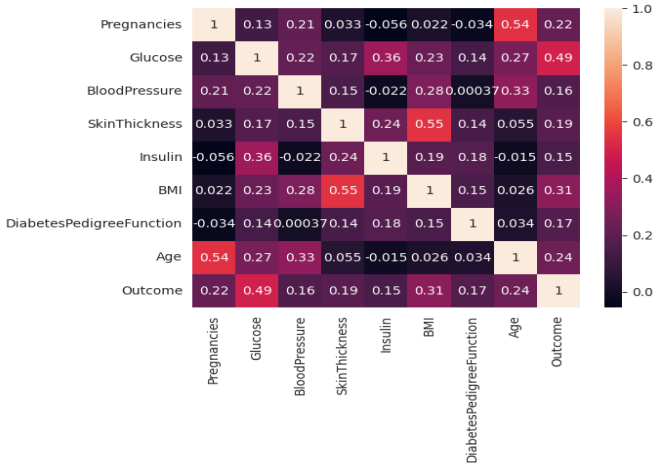

Fig. 8. Correlation score with different features

Table1

Performance comparison using Confusion Matrix

Techniques	Confusion Matrix
Logistic Regression	[93 14; 20 27]
Decision tree	[79 28; 17 30]
Random Forest	[91 16; 14 33]
XGBoost	[87 20; 14 33]
K-NN	[78 21; 18 37]
Weighted SVM	[67 32; 8 47]
FNN	[94 13; 15 32]

As per the above table, the correlated variables are considered and developed the model in order to predict the outcome variable.

The primary criterion for evaluating the performance of a machine learning model is its accuracy, which represents the proportion of correctly classified instances out of the total instances. Among all the models evaluated, the Feed Forward Neural Network (FNN) achieved the highest accuracy score of 82%. This indicates that the FNN model was able to predict diabetes with the highest level of accuracy compared to the other models.

Neural networks, especially feed-forward neural networks, are known for their ability to capture complex nonlinear relationships in the data. Unlike simpler models like Logistic Regression and Decision Trees, FNNs can handle more intricate data patterns, making them more flexible and adaptable to different types of data. This flexibility likely contributed to the superior performance of the FNN in this study, shown as Figure 9.

Table 2

Techniques with its Performance Metrics

Techniques	Accuracy	Precision	Recall	F1-Score
LR	0.78	0.6585	0.5745	0.6136
DT	0.71	0.5172	0.6383	0.5714
RF	0.81	0.6735	0.7021	0.6875
XGBoost	0.78	0.6226	0.7021	0.6600
KNN	0.75	0.6379	0.6727	0.6549
Weighted SVM	0.74	0.5949	0.8545	0.7015
FNN	0.82	0.7111	0.6809	0.6957

Ensemble Methods and Regularization

Random Forest and XGBoost are ensemble methods that combine multiple weak learners to improve prediction accuracy and generalization. While these ensemble methods performed well with accuracy of 81% and 78% respectively, the FNN outperformed them with an accuracy of 82%. Moreover, neural networks inherently include regularization techniques, such as dropout and weight decay, which help prevent overfitting and improve the model's generalization capability.

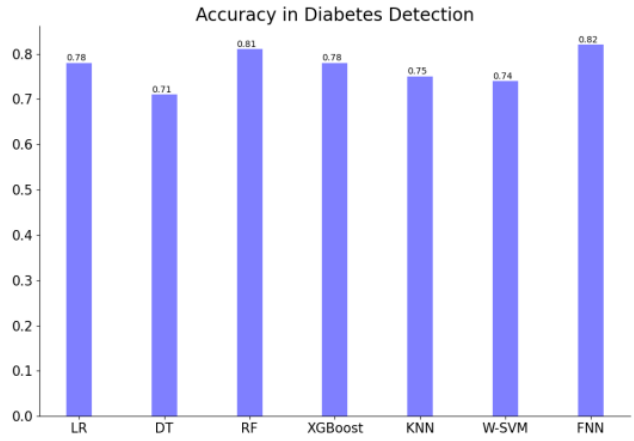

Fig. 9. Accuracy measures of different models

Comparison with Other Models

The other models, including Logistic Regression, Decision Tree, k-Nearest Neighbors, and Weighted Support Vector Machine, achieved accuracies ranging from 71% to 78%. Although some of these models performed reasonably well, they were outperformed by the FNN in terms of accuracy.

5. Conclusions

This paper represents the detailed view of predictive modelling and is in-corporate with data analytics to analyse the biological and demographic data sets of type2 diabetes patients. As per the above papers survey, some research gaps has to be addressed, that are the usage of larger datasets, outlier or abnormalities detection, improving the prediction model, can combine the optimization technique with hybrid model, application can be developed to identify the disease in the android, and usage of datasets of multiple classes.

Based on the comparative analysis of the machine learning models for predicting diabetes, the Feed Forward Neural Network (FNN) emerged as the most favorable model with an accuracy of 82%. The FNN demonstrated superior performance due to its ability to capture complex relationships in the data, inherent regularization techniques, and higher accuracy compared to other models. Therefore, for predicting diabetes using the given dataset, the FNN is recommended as the preferred machine learning model.

Author contributions

Leelambika KV 1: Conceptualization, Methodology, Software, Field study, Data curation, Writing-Original draft preparation, Software, Validation **Dr. Shanmugarathinam G 2:** Visualization, Investigation, Writing-Reviewing and Editing.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Haixia Shang and Zhi-Ping Liu, “Prioritizing Type2 Diabetes Genes by Weighted PageRank on Bilayer Heterogeneous Networks”, IEEE/ACM Transactions on Computational Biology and Bioinformatic, Vol. 18, No.1, January 2021.

[2] Liying Zhang, Yikang Wang, Miaomiao Niu, Chongjian Wang, And Zhenfei Wang, “Nonlaboratory-Based Risk Assessment Model For Type2 Diabetes Mellitus Screening in Chinese Rural Population: A Joint Bagging-Boosting Model”, IEEE Journal of Biomedical and Health Informatics, Vol. 25, No. 10, October 2021.

[3] Nada Y. Philip, Manzoor Razaak, John Chang, Suchetha. M, Maurice O’kane, & Barbara K. Pierscionek, “A Data Analytics Suite for Exploratory Predictive, and Visual Analysis of Type2 Diabetes”, IEEE Access, Volume 10, January 2022, IF: 3.467 (2021).

[4] Benjamin Lobo , Leon Farhy, Mahdi Shafiei, and Boris Kovatchev, “A Data-Driven Approach to Classifying Daily Continuous Glucose Monitoring (CGM) Time Series”, IEEE Transactions on Biomedical Engineering, Vol. 69, No.2, February 2022.

[5] Neha Prerna Tiggaa and Shruti Garg, “Prediction of Type 2 Diabetes using Machine Learning Classification Methods”, Elsevier, 2020.

[6] Harleen Kaur and Vinita Kumari, “Predictive modelling and analytics for diabetes using a machine learning approach”, Applied Computing and Informatics Vol. 18 No. 1/2, 2022 pp. 90-100,

[7] Emerald Publishing Limited, e-ISSN: 2210-8327, p-ISSN: 2634-1964, DOI 10.1016/j.aci.2018.12.004.

[8] Shamreen Ahamed, Meenakshi Sumeet Arya and Auxilia Osvin Nancy V,”Prediction of Type-2 Diabetes Mellitus Disease using machine learning classifiers and techniques, Frontiers in Computer Science, May 2022,| Volume 4,| Article 835242.

[9] Victor Chang, Jozeene Bailey, Qianwen Ariel Xu2 & Zhili Sun, “Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms”, Neural Computing and Applications (2023) 35:16157–16173

[10] Waleed Noori Hussein, Zainab Muzahim Mohammed, Amani Naama Mohammed, Identifying risk factors associated with type 2 diabetes based on data analysis”, Elsevier: Measurement: Sensors, 29 October 2022 , 2665-9174/© 2022.

[11] Sukhpreet Kaur, Yogesh Kumar, Apeksha Koul, Sushil Kumar Kamboj, “A Systematic Review on Metaheuristic Optimization Techniques for Feature Selections in Disease Diagnosis: Open Issues and Challenges”, Computational Methods in Engineering (2023) 30:1863–1895, Springer Journal.

[12] Ganeshsree Selvachandran, Shio Gai Quek, Raveendran Paramesran, Weiping Ding, Le Hoang Son, “Developments in the detection of diabetic retinopathy: a state-of-the-art review of computer-aided diagnosis and machine learning methods”, Springer Journal - Artifcial Intelligence Review (2023) 56:915–964.

[13] Sarah Shafqat, et al, “Leveraging Deep Learning for Designing Healthcare Analytics Heuristic for Diagnostics”, Neural Processing Letters (2023) 55:53–79, https://doi.org/10.1007/s11063-021-10425-w.

[14] Rashi Rastogi, Mamta Bansal, “Diabetes prediction model using data mining techniques”, Elsevier - Measurement: Sensors, 5 December 2022, 2665-9174/© 2022, https://doi.org/10.1016/j.measen.2022.100605

[15] Sumeet Kalia , Olli Saarela, Tao Chen, Braden O’Neill, Christopher Meaney, Jessica Gronsbell, “Marginal Structural Models Using Calibrated Weights With SuperLearner: Application to TypeII Diabetes Cohort”, IEEE Journal of Biomedical and Health Informatics, Vol.26, No.8, August 2022.

[16] L. B. Balzer and M. L. Petersen, “Invited commentary: Machine learning in causal inference–how do i love thee? Let me count the ways,” Amer. J. Epidemiol., vol. 190, no. 8, pp. 1483–1487, 2021.

[17] L. Lin, M. Sperrin, D. A. Jenkins, G. P. Martin, and N. Peek, “A scoping review of causal methods enabling predictions under hypothetical interventions,” Diagn. Prognostic Res., vol. 5, no. 1, pp. 1–16, 2021.

[18] Harleen Kaur, Vinita Kumari, “Predictive modelling and analytics for diabetes using machine learning approach”, vol.10, 2018

[19] Arfan Ahmed1, Sarah Aziz1, Alaa Abd-alrazaq, Faisal Farooq, Javaid Sheikh, “Overview of Artificial Intelligence–Driven Wearable Devices for Diabetes: Scoping Review, Journal Of Medical Internet Research, vol.24, 2022.

[20] Faezeh Marzbanrad, Ahsan H. Khandoker, Brett D. Hambly, “Methodological Comparisons of Heart Rate Variability Analysis in Patients With Type 2 Diabetes and Angiotensin Converting Enzyme Polymorphism”, IEEE Journal Of Biomedical And Health Informatics, Vol. 20, No. 1, January 2016.

[21] Konstantia Zarkogianni, Maria Athanasiou, Anastasia C. Thanopoulou, and Konstantina S. Nikita,“Comparison of Machine Learning Approaches Toward Assessing the Risk of Developing
Cardiovascular Disease as a Long-Term Diabetes Complication”, IEEE Journal Of Biomedical And Health Informatics, VOL. 22, NO. 5, SEPTEMBER 2018.

[22] Michele Bernardini , Luca Romeo , Paolo Misericordia, and Emanuele Frontoni ,” Discovering the Type 2 Diabetes in Electronic Health Records Using the Sparse Balanced Support Vector Machine”, IEEE Journal Of Biomedical And Health Informatics, Vol. 24, No. 1, January 2020.

[23] Lambert JR, Perumal E (2022) Oppositional frefy optimization based optimal feature selection in chronic kidney disease classifcation using deep neural network. J Ambient Intell Humaniz Comput 13(4):1799–1810

[24] Hajiha shemi V, Hassani Z, Dehmajnoonie I, Borna K (2019), Hybrid algorithms of whale optimization algorithm and k-nearest neighbor to predict the liver disease. EAI Endorsed Trans Con

[25] text-Aware Syst Appl 6(16):156838.

https://doi.org/10.4108/eai.13-7-2018.156838

[26] Gautam R, Kaur P, Sharma M (2019) A comprehensive review on nature inspired computing algorithms for the diagnosis of chronic disorders in human beings. Prog Artif Intell 8(4):401– 424. https://doi.org/10.1007/s13748-019-00191-1

[27] Kaur P, Sharma M (2019) Diagnosis of human psychological disorders using supervised learning and nature-inspired computing techniques: a meta-analysis. J Med Syst. https://doi.org/

[28] 10.1007/s10916-019-1341-2.

[29] Sharma M, Kaur P (2020) A comprehensive analysis of nature inspired meta-heuristic techniques for feature selection problem. Arch Computation Methods Eng. https://doi.org/10.1007/ s11831-020-09412-6

[30] Chander S, Padmanabha V, Mani J (2021) Jaya spider monkey optimization-driven deep convolutional LSTM for the prediction of COVID'19. Bio-Algorithms Med-Syst. https://doi.org/10.1515/bams-2020-0030.