Precision Prognosis: Leveraging Blood Biomarkers for Early Detection of Diabetes Using Machine Learning Algorithms

Robinson Victor
School of Computing, Engineering and Digital Technology
Teesside University
United Kingdom
D3739635

***Abstract*—Diabetes can lead to several complications, such as heart disease, stroke, blood vessel damage, retinopathy or blindness, and nephropathy. This condition has no cure, so early detection is vital to manage and reduce the complications associated with it. For the early detection of diabetes, predictive prognosis using blood biomarkers and machine learning has gained significant attention. Machine learning has evolved as a valuable tool for predictive modeling and is widely used in healthcare because of its ability to identify patterns that may elude traditional experimental approaches. In this study, we utilized data comprising medical information and laboratory analyses of patients from Medical City Hospital and the Specialized Center for Endocrinology and Diabetes at Al-Kindy Teaching Hospital, Iraq. Three binary classification models were developed using nine different blood biomarkers in combination with various physical factors, such as age, gender, and body weight, known to influence this disease. These factors were integrated with different machine learning algorithms, including logistic regression, K-nearest neighbor classifier, and decision tree. The models were evaluated using a range of performance metrics, such as accuracy, recall, precision, F1-score, and ROC curve. All three models performed well in predicting diabetes, with decision tree emerging as the top performer among them. It displayed a better average score on 10-fold cross-validation and outperformed others on both training and testing data, achieving accuracy scores of 99% and 96%, respectively. These findings hold promise for the prediction of diabetes in patients.**

Keywords—Diabetes, Logistic Regression, KNN, Decision Tree, Predictive Modeling, Cross-validation, Machine Learning

I. Introduction

Diabetes, scientifically known as diabetes mellitus, is a medical condition that arises from the body's inability to effectively regulate blood glucose levels [3], which is a type of blood sugar used as fuel to power the body [4]. This condition is often characterized by a high concentration of glucose in the blood over a long period of time [7]. Diabetes often results from difficulties in the body cells ability to effectively absorb and store sugar, typically due to dysfunction of the pancreas. This dysfunction impairs the production and secretion of sufficient and effective insulin to support metabolism [3]. Insulin is the hormone responsible for stimulating the body cells to absorb and store glucose to be used as a source of energy [28].

Diabetes is a growing concern worldwide, as a recent study by [25] shows that over half a billion people all over the world have diabetes, accounting for over 10.5% of the world's adult population aged between 20 and 79 years old. The prevalence is most dominant between 75 and 79 years old, with projections indicating it rising to 12.2%, encompassing about 783.2 million individuals in the years to come. Reference [25] further stated that money spent worldwide on cases related to diabetes was approximately 966 billion USD in 2021 and may reach 1 trillion USD by 2045.

Several factors, such as age [25], obesity [3], gender, physical inactivity, poor diet, and genetics [19] can also induce diabetes [20]. Diabetes leads to a growing number of complications such as retinopathy, nephropathy, neuropathy [21], vascular disease [17], tuberculosis, slow healing leading to amputation of the limbs due to ulcers [19], reduced life expectancy [3], and even death. Among the several complications of diabetes, it is also generally seen as a disability under the UK Equality Act 2010 and the Disability Discrimination Act 1995 [4]. While there is currently no cure for diabetes, early detection can significantly reduce the complications caused by this disease [30]. Traditional diabetes testing methods often rely on in vivo or in vitro observations, which may not always be adequate for early detection. However, the intersection of AI/ML and healthcare has shown promise in addressing these challenges by swiftly identifying patterns that may outpace traditional methods.

Reference [8], [7], [20], [28], [21], [19], [15] and [16] utilized several ML algorithms to develop models for classifying diabetes in patients. However, while their studies yielded results, they primarily used the most essential biomarker, blood glucose, for their predictions. Overall, a person's blood glucose level can provide insight into their risk of diabetes, which either alters their study or diminishes its importance. Reference [14] and [9] also developed predictive models for classifying diabetes. They achieved results, but the models included glycated hemoglobin (HbA1c), which is closely related to blood glucose and same is the case as the previous.

This study aims to leverage indirectly linked blood biomarkers associated with diabetes, such as blood urea nitrogen, creatinine ratio, lipids, and lipoproteins [27], alongside body mass index, to identify patterns and facilitate early diabetes prediction. Some of these biomarkers may be byproducts of various conditions that may be induced by diabetes but not directly linked to it. For instance, creatinine, a byproduct filtered from the body by the kidneys, results from muscle breakdown. This breakdown may be due to factors such as rapid weight loss or kidney issues which may be induced by diabetes, but creatinine levels are also associated with various other health conditions.

The experiment was developed using three diverse ML algorithms, such as logistic regression, K-nearest neighbor classifier, and decision tree. These models were chosen for their simplicity and ease of implementation. KNN is a non-parametric classifier that predicts based on the proximity of the data [10], enabling it to handle multiclass. On the other hand, decision tree is a tree-based non-parametric classifier known for its ability to handle imbalanced data, and manage missing values, and also logistic regression, for its interpretability. Performance metrics were computed for each model, with decision tree outperforming all other models on both train and test data with 99% and 96% accuracy, respectively. The study uses data consisting of various medical records of patients from the Specialized Center for Endocrinology and Diabetes at Al-Kindy Teaching Hospital, Iraq which can be accessed here.

II. Methodology

This section entails methods utilized for this study, which are highlighted in the subsections below.

A. *Structure Investigation*

Several methods were utilized to examine the overall structure, shape of the dataset, and the type of data in the features. This revealed that the dataset contains a total of 1007 samples and 14 features, of which 2 are non-numerical features and 12 are numerical features. The *describe* method was specifically useful in generating statistical summaries.

B. *Quality Investigation and Data Preprocessing*

The dataset contained 162 duplicate entries, which were removed, bringing the number of rows to 845. Null checks revealed no missing values. Using Pandas plot, it was shown that some numerical features contained a few unwanted entries. Some biomarkers exhibited unusually high values. Further investigation revealed that while some of these values were biologically implausible, others were justifiable given the nature of the medical condition, which can contain elevated biomarker levels. As a result, it was concluded that there was no need to further check for outliers.

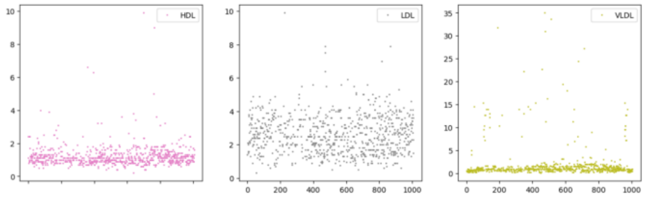

Fig. 1. Pandas plot showing some lipoproteins with errors

C. *Exploratory Data Analysis and Content Investigation*

Further examination, utilizing appropriate visualizations, aims to deepen understanding of the features and their relationship with the target variable which may not be seen normally. This process can help to uncover various insights including more about the imbalanced nature of the dataset.

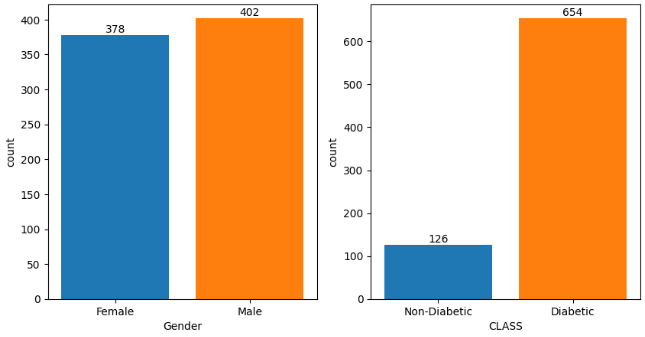

Fig. 2. Distribution of Gender feature and target CLASS

Fig. 2 shows the distribution of gender and the target feature CLASS, revealing that Male make up 51.5% (402) of the sample, while Female make up 48.5% (378), and also highlighting an imbalance in the label variable.

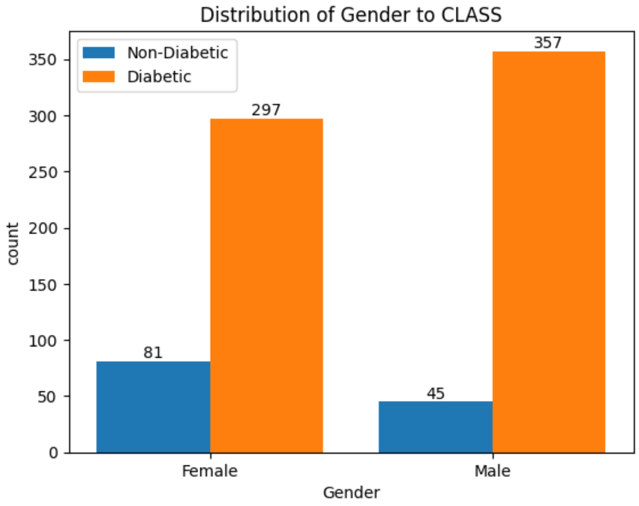

Fig. 3. Distribution of Gender to the target

Fig. 3 shows the distribution of Gender feature to the label variable. From this, 88.8% (357 out of 402) of Male patients and 78.6% (297 out of 378) Female patients have been diagnosed with diabetes.

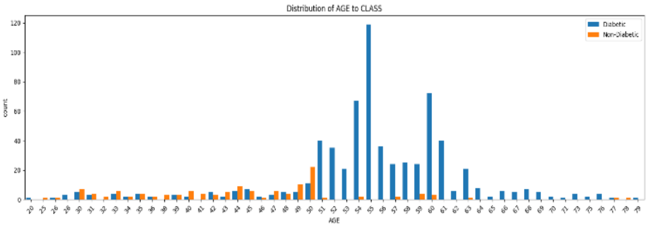

Fig. 4. Distribution of AGE to the target

Fig. 4 illustrates the distribution of AGE to CLASS variable. Age is a critical factor in diabetes, and the figure indicates that higher age correlates with an increased risk of developing diabetes.

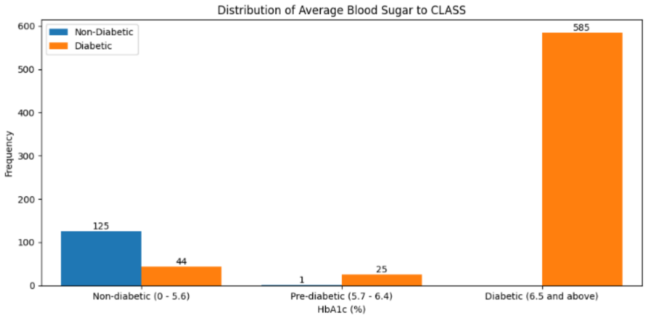

Fig. 5. Distribution of HbA1c (average blood sugar) to the target

Fig. 5 shows a clear relationship between HbA1c and the target. Average blood sugar, being the most important biomarker for detecting diabetes [29], provides insight into a person's risk of developing the disease. This feature was not explicitly included in the model's development.

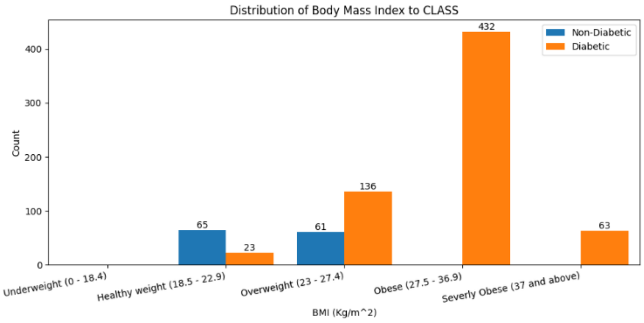

Fig. 6. Distribution of BMI to the target variable

Fig. 6 clearly shows that people in the overweight category and above have higher risk of developing the disease.

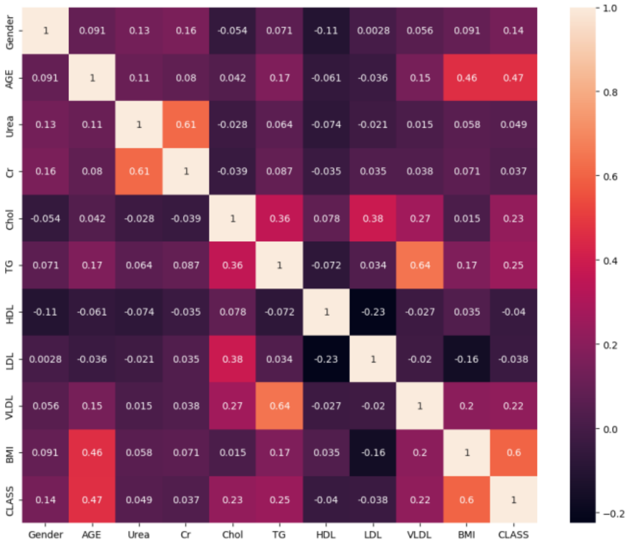

Fig. 7. Correlation matrix

Fig. 7 shows not too high positive correlation between featrues and the target.

D. *Feature Engineering*

It is crucial to remove unimportant features during data preparation. **'ID'** and **'No_Pation**' were dropped because they served no real purpose and were only used as unique identifiers. '**HbA1c**' was also dropped because it is closely related to the target; in fact, the average blood sugar alone is sufficient to determine whether a person has diabetes.

1) Data Encoding: There were only two categorical features in the dataset and both were correctly encoded [10].

2) Feature Selection: Only a subset of the most relevant features were selected [11] for the model development. There weren't any high multicollinearity.

E. *Model Development*

This section entails steps taken to train the model.

1) Train-test split: The target feature is separated from the predictor variables. The dataset is then divided into two parts, 80% (0.8) for training the model and 20% (0.2) for testing its performance [22].

2) Feature Scaling: Scaling is a critical part of ML, MinMaxScaler was used to adjust the values of features to have a standard scale of measure.

3) Fitting the Model: Three diverse ML algorithms were used for the development of the model which are, Logistic Regression, K-nearest neighbor, and Decision tree. All three are popular baseline supervised learning techniques used for classification [22] and were selected for their simplicity and ease of implementation. The final is a binary classification model that involves classifying samples into one of two classes, predicting the risk of diabetes in a patient.

a) Logistic Regression is great for its interpretability.

b) KNN is non-parametric and predicts based on the nearest neighbor in the data.

c) Decision Tree is tree-based, non-parametric and is best for its ability to handle imbalanced data, and manage missing values.

F. *Model Evaluation*

In this section, the performance of the model is evaluated using various metrics such as accuracy, precision, recall, and F1-score, computed from the four components of the confusion matrix (TP, TN, FP, and FN). The evaluation is done using the test set and it aims to assess the effectiveness of the model in making accurate predictions.

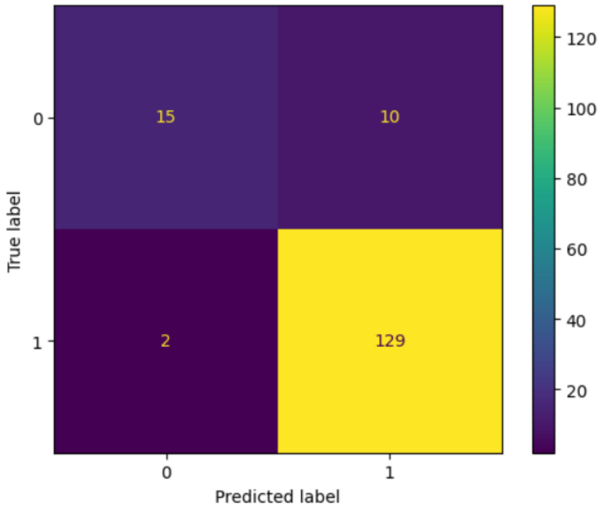

Fig. 8. Confusion Matrix for Logistic Regression

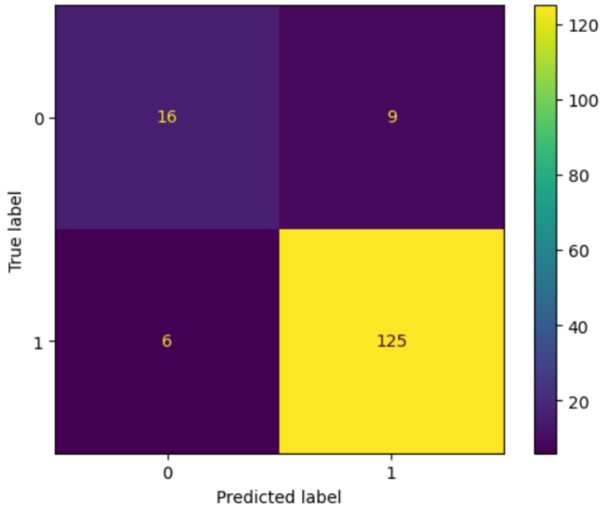

Fig. 9. Confusion Matrix for KNN

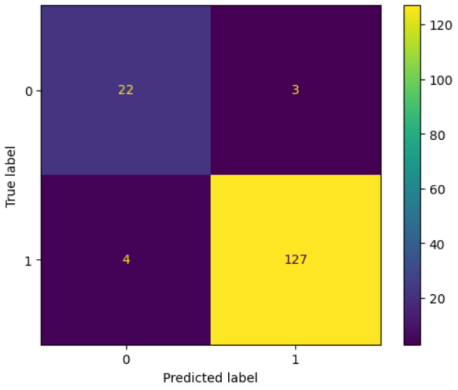

Fig. 10. Confusion matrix for Decision Tree

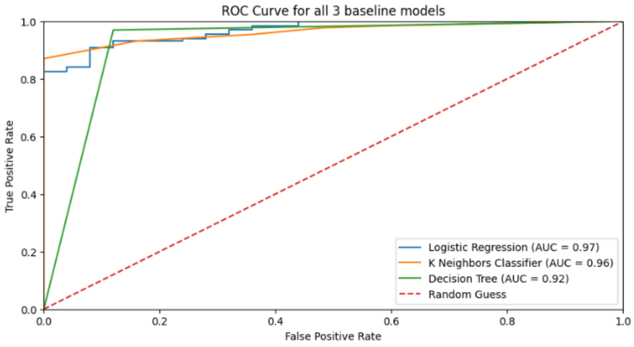

Fig. 11. ROC curve of the model

TABLE I. MODEL PERFORMANCE METRICS

Model	*Train Accuracy*	*Accuracy*	*Recall*	*Precision*	*F1-Score*
LG	0.947	0.923	0.984	0.928	0.92
KNN	0.958	0.903	0.954	0.932	0.90
DT	1.000	0.955	0.969	0.976	0.96

a. Model performance metrics.

TABLE I. shows DT is the highest performing model, but because the dataset is imbalanced and the model hasn't been tuned yet, the ROC AUC will be the crucial metric. Fig. 11 shows LR achieved an ROC AUC score of 0.97, making it the best-performing algorithm for the baseline model.

G. *Model Optimization*

Here steps are taken to improve the models' performance.

1) SMOTE: Fig. 2 shows the imbalanced nature of the dataset. Synthetic minority oversampling technique [18] was utilized to balance the number of underrepresented and dominant groups in the train set.

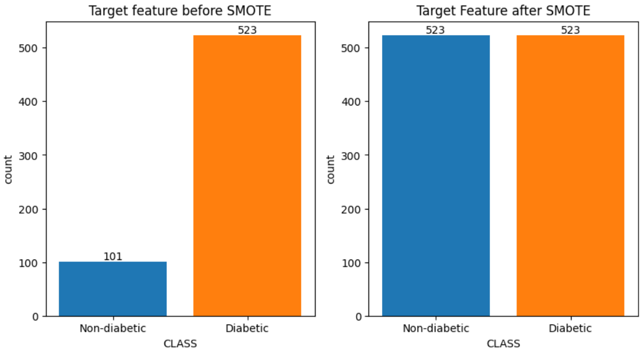

Fig. 12. Distribution of CLASS before and after SMOTE

2) K-fold Cross-validation: The purpose of this is to evaluate the model with non-overlapping test sets [10] to achieve better generalization to unseen data. As suggested in [11], 10-fold cross-validation was performed, and the average estimate was computed. This helped reduce variance, resulting in an improvement in the overall accuracy score.

3) Hyperparameter Tuning: Hyperparameters are a set of values specified to influence the performance of a model. The GridSearch hyperparameter tuning technique was employed to identify the best combination of hyperparameters that yield the best performance. This technique conducts an exhaustive search within the defined search space, ensuring that the best values are not missed. As a result, the overall performance of the model was enhanced

TABLE II. METRICS OF THE TUNED MODEL

Models	*Train Accuracy*	*Test Accuracy*
Logistic Regression	0.966520	0.923077
K-Nearest Neighbor	0.982766	0.935897
Decision Tree	0.989505	0.955128

Tuned model performance metrics.

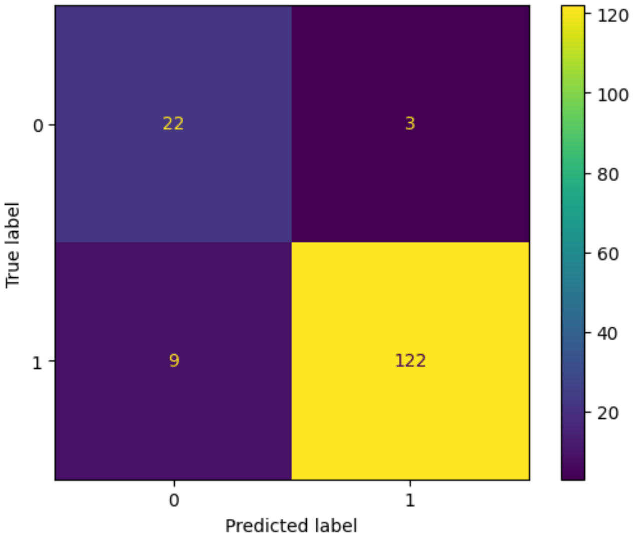

Fig. 13. Confusion matrix of the best Logistic Regression model

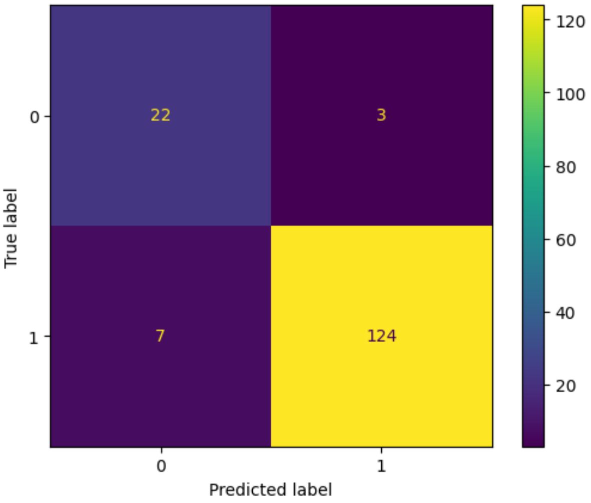

Fig. 14. Confusion matrix of the best KNN model

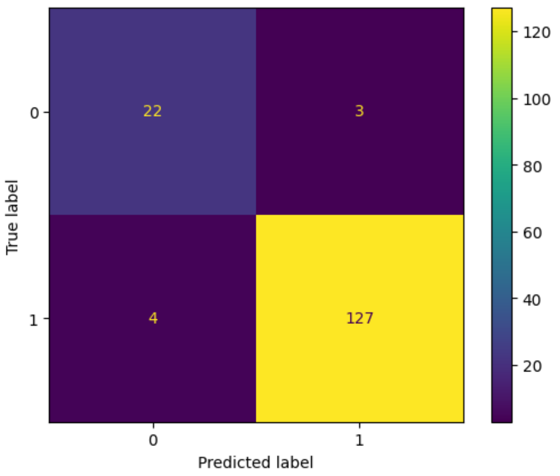

Fig. 15. Confusion matrix of the best Decision Tree model

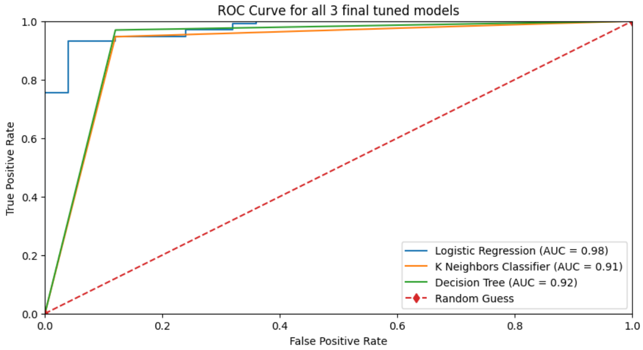

Fig. 16. ROC Curve for the fnal tuned models

H. Model Explanation

The essence of this is to see which features contribute the most to the model outcome or performance.

1) Feature Importance:

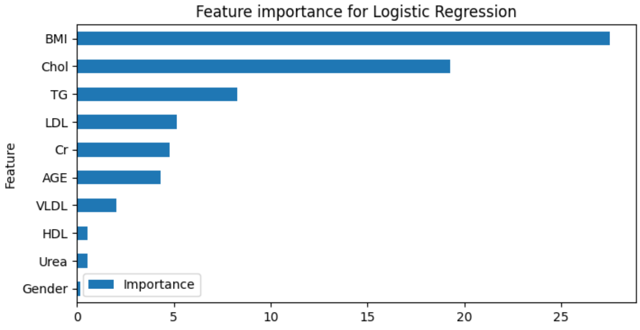

Fig. 17. Feature importance for Logistic Regression

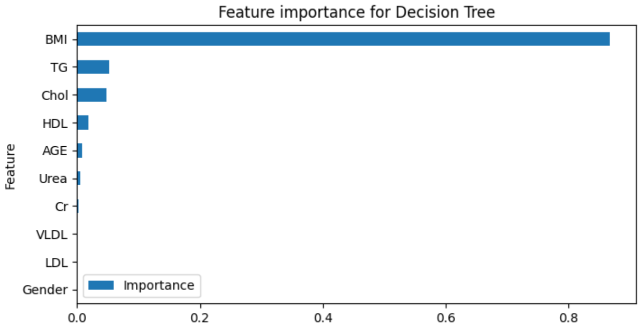

Fig. 18. Feature importance for Decision Tree

2) SHAP:

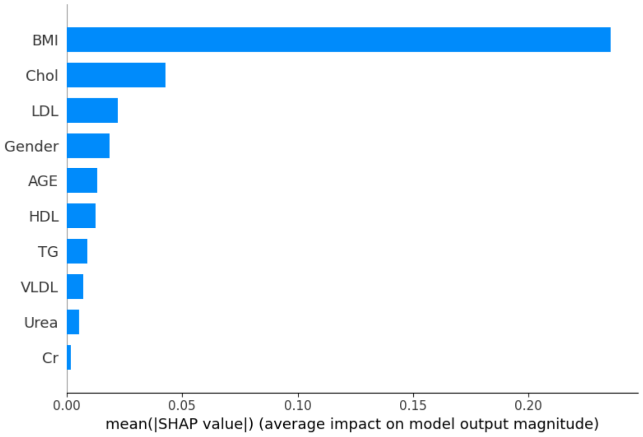

Fig. 19. SHAP plot for KNN

III. Results and Discussion

Upon examining all major performance metrics for the final tuned models, the comparative analysis in TABLE III. suggests that Decision Tree outperforms all other models. This differs from the findings in Fig. 11, where logistic regression emerged as the top performer based on ROC AUC for the baseline model, prior to balancing the data and optimizing the models.

TABLE III. METRICS OF THE FINAL TUNED MODELS

Model	*Train Accuracy*	*Accuracy*	*Recall*	*Precision*	*F1-Score*
LG	0.966	0.923	0.931	0.976	0.92
KNN	0.982	0.935	0.946	0.976	0.94
DT	0.989	0.955	0.969	0.976	0.96

IV. Future work and Conclusion

This project yielded exciting results; however, there were various limitations. From the dataset source, the units for biomarkers were not specified, requiring additional time and effort to carry out research and compare figures in order to identify the appropriate units for each biomarker. Also, while it is ideal to perform 10-fold cross-validation 10 times [11], it was computationally expensive to do so with grid search while running the code on a local machine. Even after setting the “n_jobs” parameter to -1 to utilize all CPUs, it was still slow. Additionally, from the feature importance analysis in [Model Explanation] section, BMI is shown to be a crucial feature for all three models, contributing the most to the model outcomes. However, this somewhat contradicts with the essence of the project, which aims to develop a model with indirectly linked traits; the reason "HbA1c" was not included. Nevertheless, it's reassuring to note that, in reality, BMI primarily reflects body weight and doesn't definitively indicate diabetes.

Future work should involve gathering more information and exploring additional novel biomarkers indirectly linked to the disease. Also, using a wider range of machine learning algorithms and comparing results would be beneficial.

In conclusion, this project aimed to develop a predictive model for the early detection of diabetes using indirectly linked blood biomarkers. ML has proven to be a valuable tool, and thus, this approach will potentially save more lives.

References

[1] Eyth E, Naik R. Hemoglobin A1C. [Updated 2023 Mar 13]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available at: https://www.ncbi.nlm.nih.gov/books/NBK549816/

[2] Hosten AO. BUN and Creatinine. In: Walker HK, Hall WD, Hurst JW, editors. Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition. Boston: Butterworths; 1990. Chapter 193. Available from: https://www.ncbi.nlm.nih.gov/books/NBK305/

[3] Nakrani MN, Wineland RH, Anjum F. Physiology, Glucose Metabolism. [Updated 2023 Jul 17]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK560599/

[4] Hantzidiamantis PJ, Awosika AO, Lappin SL. Physiology, Glucose. [Updated 2022 Sep 19]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK545201/

[5] Chapple, B. (2024) Your legal rights when you have diabetes, Diabetes UK. Available at: https://www.diabetes.org.uk/guide-to-diabetes/life-with-diabetes/your-legal-rights (Accessed: 06 May 2024).

[6] Abcar, A.C., Chan, L. and Yeoh, H. (2004) 'What To Do for the Patient with Minimally Elevated Creatinine Level?', The Permanente Journal, 8(1), pp. 51-53. Available at: https://doi.org/10.7812/TPP/03-119

[7] Alaa Khaleel, F. and Al-Bakry, A.M. (2023) 'Diagnosis of diabetes using machine learning algorithms', Materials Today: Proceedings, 80, pp. 3200-3203. Available at: https://doi.org/10.1016/j.matpr.2021.07.196

[8] Butt, U.M. et al. (2021) 'Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications', Journal of Healthcare Engineering, 2021, pp. 9930985. Available at: https://doi.org/10.1155/2021/9930985

[9] Cahn, A. et al. (2020) 'Prediction of progression from pre-diabetes to diabetes: Development and validation of a machine learning model', Diabetes/Metabolism Research and Reviews, 36(2), pp. e3252. Available at: https://doi.org/10.1002/dmrr.3252

[10] Campesato, O. (2020a) '4. Intro to Machine Learning', in '4. Intro to Machine Learning', Angular and Machine Learning Pocket Primer. Berlin, Boston: Mercury Learning and Information, pp. 135-172.

[11] Campesato, O. (2020b) 'Chapter 7 Natural Language Processing and Reinforcement Learning', in 'Chapter 7 Natural Language Processing and Reinforcement Learning', Python 3 for Machine Learning. Berlin, Boston: Mercury Learning and Information, pp. 209-234.

[12] Chow, L.S. et al. (2016) 'Biomarkers associated with severe hypoglycaemia and death in ACCORD', Diabetic Medicine, 33(8), pp. 1076-1083. Available at: https://doi.org/10.1111/dme.12883

[13] Chung, W.K. et al. (2020) 'Precision medicine in diabetes: a Consensus Report from the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD)', Diabetologia, 63(9) Available at: https://doi.org/10.1007/s00125-020-05181-w

[14] Dagliati, A. et al. (2018) 'Machine Learning Methods to Predict Diabetes Complications', Journal of Diabetes Science and Technology, 12(2), pp. 295-302. Available at: https://doi.org/10.1177/1932296817706375

[15] Kodama, S. et al. (2022) 'Predictive ability of current machine learning algorithms for type 2 diabetes mellitus: A meta-analysis', Journal of Diabetes Investigation, 13(5), pp. 900-908. Available at: https://doi.org/10.1111/jdi.13736

[16] Lai, H. et al. (2019) 'Predictive models for diabetes mellitus using machine learning techniques', BMC Endocrine Disorders, 19(1), pp. 101-6. Available at: https://doi.org/10.1186/s12902-019-0436-6

[17] Lansang, M.C. et al. (2022) Diabetes : Clinician's Desk Reference. Milton: Taylor & Francis Group.

[18] Lema ^itre, G., Nogueira, F. and Aridas, C.K. (2017) 'Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning', Journal of Machine Learning Research, 18(17), pp. 1-5. Available at: http://jmlr.org/papers/v18/16-365.html

[19] Md Shahin Ali et al. (2023) 'A Novel Approach for Best Parameters Selection and Feature Engineering to Analyze and Detect Diabetes: Machine Learning Insights', BioMed Research International, 2023 Available at: https://doi.org/10.1155/2023/8583210

[20] Mujumdar, A. and Vaidehi, V. (2019) 'Diabetes Prediction using Machine Learning Algorithms', Procedia Computer Science, 165, pp. 292-299. Available at: https://doi.org/10.1016/j.procs.2020.01.047

[21] Nicolucci, A. et al. (2022) 'Prediction of complications of type 2 Diabetes: A Machine learning approach', Diabetes Research and Clinical Practice, 190, pp. 110013. Available at: https://doi.org/10.1016/j.diabres.2022.110013

[22] Pedregosa, F. et al. (2011) 'Scikit-learn: Machine Learning in Python', Journal of Machine Learning Research, 12(85), pp. 2825-2830. Available at: http://jmlr.org/papers/v12/pedregosa11a.html

[23] Pigeyre, M. et al. (2022) 'Identifying Blood Biomarkers for Type 2 Diabetes Subtyping: A Report From the ORIGIN Trial', Canadian Journal of Diabetes, 46(7, Supplement), pp. S5. Available at: https://doi.org/10.1016/j.jcjd.2022.09.013

[24] Pundir, C.S., Kumar, P. and Jaiwal, R. (2019) 'Biosensing methods for determination of creatinine: A review', Biosensors & Bioelectronics, 126, pp. 707-724. Available at: https://doi.org/10.1016/j.bios.2018.11.031

[25] Sun, H. et al. (2022) 'IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045', Diabetes Research and Clinical Practice, 183, pp. 109119. Available at: https://doi.org/10.1016/j.diabres.2021.109119

[26] Tedre, M. and Moisseinen, N. (2014) 'Experiments in Computing: A Survey', The Scientific World Journal, 2014, pp. 549398. Available at: https://doi.org/10.1155/2014/549398

[27] Tibor V Varga et al. (2021) 'Predictive utilities of lipid traits, lipoprotein subfractions and other risk factors for incident diabetes: a machine learning approach in the Diabetes Prevention Program', BMJ Open Diabetes Research & Care, 9(1), pp. e001953. Available at: https://doi.org/10.1136/bmjdrc-2020-001953

[28] Tigga, N.P. and Garg, S. (2020) 'Prediction of Type 2 Diabetes using Machine Learning Classification Methods', Procedia Computer Science, 167, pp. 706-716. Available at: https://doi.org/10.1016/j.procs.2020.03.336

[29] Lyons TJ, Basu A. Biomarkers in diabetes: hemoglobin A1c, vascular and tissue markers. Transl Res. 2012 Apr;159(4):303-12. doi: 10.1016/j.trsl.2012.01.009. Epub 2012 Jan 31. PMID: 22424433; PMCID: PMC3339236.

[30] Waqas Khan Q, Iqbal K, Ahmad R, Rizwan A, Nawaz Khan A, Kim D. 2024. An intelligent diabetes classification and perception framework based on ensemble and deep learning method. PeerJ Computer Science 10:e1914 https://doi.org/10.7717/peerj-cs.1914