Diabetes Prediction through Machine Learning Algorithms

Mathukiya Vaibhav Jagdish
Student ID: 127837
University of Camerino
Department of computer science
vaibhavjag.mathukiya@studenti.unicam.it

Obude Destiny
Student ID: 131986
University of Camerino
Department of computer science
destiny.obude@studenti.unicam.it

Abstract - Health has always been a factor for concern, even more so given the dynamics of modern life when we consider our surroundings, lifestyle, and dietary habits. This paper presents a detailed analysis of diabetes prediction using machine learning techniques. Whereas there was research conducted by the International Diabetes Federation (IDF) back in 2021, which showed that about 10.5 percent population in the world had diabetes, this approximates around 537 million adults living with diabetes, and it has conducted a prediction which says around 1 in every 8 adults worldwide will have diabetes by 2045. This research paper aims to understand the factors that can influence diabetes and to gather relevant data. Additionally, it seeks to develop algorithms to predict the likelihood and risks of diabetes in the future and provide recommendations to prevent or manage it, if necessary. The implementation of the framework used the Pima Indians Diabetes Dataset from Kaggle. The data had been preprocessed, followed by the health indicators and feature selection with the information. The data was then extracted and used for Logistic Regression, Random Forest, Gradient Boosting, and XGBoost. This research was conducted to contribute to the field by showing the effectiveness of feature engineering, advanced machine learning algorithms to help predict diabetes, which could help medicinal strategies and personalised healthcare recommendations.

I. Introduction

Diabetes is a medical condition caused by an altered insulin response, which is characterised by hyperglycaemia, an excessive presence of blood sugar in the body. It is categorised in two stages, each of which implicitly describes its cause — Type 1 diabetes arises due to insufficient insulin production, on the other hand, type 2 diabetes arises due to ineffective use of insulin. However, pregnancy can also be a catalyst for developing diabetes.

The impact of diabetes extends beyond the individual as it can cause cardiovascular disease, kidney failure, blindness, and lower limb amputation - all very terrible consequences for the individual, which also add a burden on the healthcare system.

Diagnosing diabetes (particularly type 2 diabetes) can be challenging as there are various factors (e.g. ethnicity, BMI or other medical conditions) that could lead to the medical condition, and situations such as late diagnosis could exacerbate an individual's health condition. Additionally, patients could greatly benefit from personalised health care, designed according to their needs.

In 2025, it is estimated that over 500 million adults (20-79) are living with diabetes, and expected to increase by approximately 40% within the next two decades. Timely detection of this disease reduces the complications that would otherwise arise from it.

Machine learning techniques are increasingly relevant and vital for the detection and controlled monitoring of this disease. The research on optimising the trustworthiness of these techniques applied to diabetes prediction is still an open case, partly because there are no standards set in place across all contexts to determine what constitutes a satisfactory result. Such standards would be difficult to establish in the first place because the data is complex (biological and lifestyle changes over an extended period of time, e.g. weight gain and blood sugar trends) and sometimes incomplete (or noisy). Engineering the predictive parameters, therefore, reveals itself to be challenging.

The proliferation of wearable devices allows new perspectives and possibilities in terms of how technology intertwines with daily life and personal he-

alth. One beneficial side-effect is the facilitation of how health-related data is collected. The availability of data makes data-driven approaches more accessible in the field of healthcare research. However, new solutions should preferably easily integrate into already existing frameworks such as the electronic health record (EHR), and should be considered trustworthy by healthcare professionals. After all, healthcare professionals do require interpretable processes, which is mostly not the case in these black-box operation models. Another non-trivial issue is certainly biased data. Datasets that heavily represent a set of population may reveal data that is unsuitable for other populations.

This paper proposes a solution that pairs supervised machine learning models for classifying Pima Indian patients into at-risk or healthy status according to health indicators, to pre-emptively identify at-risk patients, with a lifestyle recommendation software that guides the patient towards healthier life choices through health-conscious suggestions.

The methodology applied is the use of an aggregation of several supervised machine learning algorithms trained on the Pima Indian women dataset to carry out diabetes predictions. Of the several models trained, the most effective one, assessed via some metric, is selected to make the prediction.

This research aims to offer a tool for the diagnosis, monitoring, and ideally, prevention of diabetes through the offering of adequate lifestyle adjustments based on patient health indicators. This would assign the individual to a central role in terms of handling their health, while assisting healthcare professionals and hospitals with efficiency.

The document is structured in a way such that, in Introduction (I) a broad overview of the research work and context is introduced. Background (II) introduces the research background, expands on the biological background of diabetes and presents the consultation of relevant literature. Methodology (III) presents the methodology, portraying all the key components that were taken into account and the process through which the work was carried out. In Discussion (IV) the implications, future possibilities and limitations of the research are discussed briefly. Finally, in Conclusions (V) the conclusions are given.

II. Background

Diabetes is a metabolic disorder characterised by the body's impaired ability to regulate blood glucose levels. This is typically due to either insufficient production of insulin (as in Type 1 diabetes) or the body's reduced sensitivity to insulin (as in Type 2 diabetes). As a result, glucose accumulates in the bloodstream, leading to a condition called hyper–glycaemia. Insulin is a hormone produced by the pancreas that facilitates the uptake of glucose by the body's cells. Chronic hyperglycaemia can damage organs and tissues over time, including the eyes, kidneys, nerves, and blood vessels. However, individual genetics plays a significant role in the likelihood of certain conditions leading to the development of the medical condition, as some genetic dispositions render individuals more susceptible than others. This is even more evident across different ethnicities. For example, Asians are two to four times more likely to develop diabetes compared to Europeans, due to their higher insulin resistance even at a normal BMI. Hispanics may develop it due to higher rates of abdominal obesity. Overall, Body fat distribution, varying responses to insulin, differences in diet, physical activity, and healthcare access all play a key role in whether the condition will be contracted or not, and these vary according to ethnicity.

All of this is important because it affects how data is collected. Different ethnicities have different risk profiles, disease progression and response to treatment. Under-representation of certain groups compared to others leads to biased models. The benefit is that personalised treatments can be tailored according to individual needs.

Literature Review

The prediction of medical conditions as a whole has been receiving quite a lot of attention from researchers, as it is a turning point in the entwinement of humans and technology. Researchers from the University of Camerino, Battineni et al., 2020, conducted a study where they reviewed the landscape of research work from 2015 to 2019 on the application of machine learning to the prediction of chronic diseases (diabetes falls in the category). In this research, over 400 research papers were reviewed, and a select few were taken aside for in-depth examination. They found out that not only is there no consensus on what the best approach is, but also that among all the produced work, the accuracy of the models ranged between 73.1 and 91.6%

Diabetes prediction is an ongoing research point in the scientific community. An extensive amount of research work has been dedicated to it, ranging from applications of classical machine learning techniques to deep learning techniques. Until a few years ago, the highest amount of accuracy achieved from models so far, according to research conducted by Larabi Marie-Sainte et al., 2019 is 95.1%, coming from deep learning techniques of CNN in combination with LSTM. Other works have been conducted using a diverse set of algorithms and techniques. The hybrid architecture proposed in the research of Larabi served as the basis for a recent research con-

ducted by Hasan and Yasmin, 2025, who introduced a model named DNet, featuring both CNN and LSTM layers. This design enables the model to effectively extract deep features through convolutional layers and capture temporal dependencies within the data via LSTM layers. Additionally, the use of residual blocks with skip connections improves information flow, while Batch Normalisation and Dropout provide regularisation to prevent overfitting. As a result, DNet achieves higher predictive performance, boasting an accuracy of 99.79% and a ROC-AUC of 99.98%, which is significantly superior in comparison to other models considered at the time of the work, the closest being proposed by Khaleel and Al-Bakry, 2021 with a Logistic Regression model of 94%. The veracity of the results is still under discussion as it has yet to be published in a peer-reviewed journal.

The most impressive models from the reviews of other works all opted for multi-model frameworks. This research somewhat follows the same philosophy in its choice of the aggregation of several models, selecting the best according to the evaluation metrics, to make the final prediction of the patient's condition.

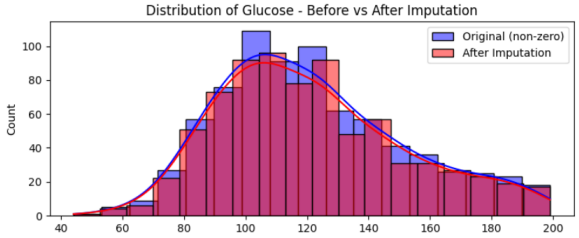
III. Methodology

The methodology of this research involved several systematic steps which led to the development of the final framework. The dataset is the Pima Indians Diabetes Dataset from Kaggle. Preprocessing operations were carried out on the data, including an oversampling to handle class imbalance. Additionally, feature augmentation was used to aid the learning capacity of the models. A collection of supervised machine learning algorithms was selected to serve as the predictive models. Predictions are given by the most suitable model based on the ROC curve value. The performance of the models is measured via several evaluation techniques, among which the ROC-AUC stands out as the selection criteria for prediction.

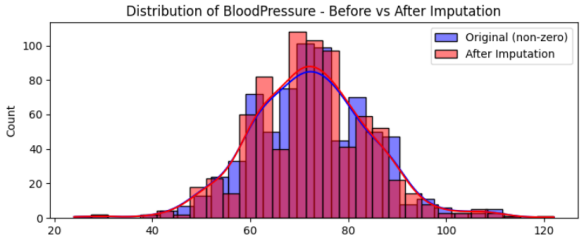
Data collection and Preprocessing

The dataset is very specific, belonging to the data of the Pima Indian Women group, and was collected based on some preconceived assumptions about what are considered relevant factors in the onset of diabetes.

Exploratory data analysis is first carried out on the data to examine its state. This helps to determine how to interact with it in terms of preparation for the task. The data undergoes a preprocessing step, where missing data are initially replaced with NaN to identify potential issues. This preprocessing allowed the recognition of physiologically impossible zero values in the key features:
Glucose: 5 zeros (0.65%), BloodPressure: 35, Insulin: 374 zeros (374 zeros (48.70%) and BMI: 11 zeros (1.43%), then these values were substituted with imputed values using the IterativeImputer from scikit-learn. This technique estimates missing values based on relationships between features using regression models. The figures 1a and 1b are some examples of the results of imputation.



(a) Skin Thickness imputation



(b) Insulin imputation

Figure 1: Comparison of imputation for Skin Thickness and Insulin features

Data imbalance is handled via oversampling to compensate for the under-representation of certain features. The original database contained 500 non-diabetic cases and 268 diabetic cases. The Synthetic Minority Over-Sampling Technique (**SMOTE**) was used to balance the classes which resulting in 500 instances in each class. The balanced dataset was then split between a training and testing set, maintaining the original distribution of the dataset. Categorical features are also adequately normalised.

Feature selection

Feature selection and augmentation are applied; the most relevant features are selected (correlated ones, therefore less relevant, are dropped). The correlation between the features can be seen in Figure 2. The remaining ones are then combined and transformed in various ways to generate new, valuable features that express meaningful relationships. To that end, twenty new features were created. The performed feature selection served to reduce dimensionality and avoid multicollinearity. Mutual information analysis was applied to select the most informative features for prediction. Figure 3 presents an overview of feature importance. This allowed the

data to be understood correctly and categorised according to the desired results or outcomes. While the feature selection was ongoing, we also unravelled potential insights from how the inclusion of certain features could affect the results.

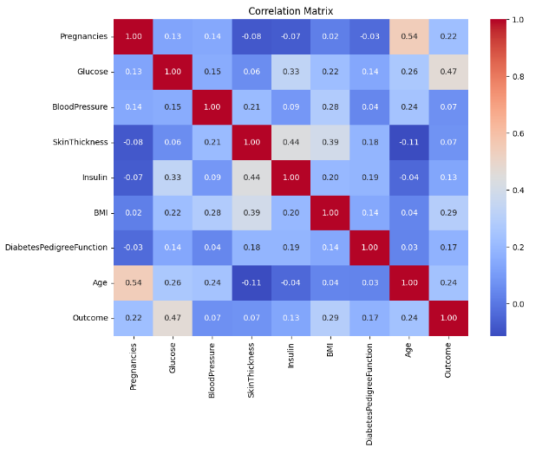


Figure 2: Correlation matrix

The following is a list of such features that were taken in consideration during the feature selection:

- **Interaction Terms**
  - `Age_BMI_Interaction`: This captures how age might modify the effect of BMI on diabetes risk.
  - `Glucose_BMI_Interaction`: This shows the combined effect of glucose levels and BMI.
  - `Glucose_BMI_Age_Interaction`: This is a three-way interaction term for this.
- **Ratio Features**
  - `Glucose_to_Insulin_Ratio`: Important for assessing insulin resistance.
  - `BMI_to_Age_Ratio`: Checks BMI with respect to age.
  - `Glucose_to_BMI_Ratio`: Normalises glucose levels by body mass.
- **Logarithmic Transformations**
  - `Insulin_log`
  - `BMI_log`
  - `HOMA_IR`: Homeostatic Model assessment for Insulin resistance.
- **Categorical features**
  - `BMI categories`: `BMI_Category_Obese_I`, `BMI_Category_Obese_II_III`, `BMI_Category_Overweight`, `BMI_Category_Underweight`, `BMI_Category_Normal`
  - `Age groups`: `Age_Group_Young`, `Age_Group_Middle_Aged`, `Age_Group_Senior`, `Age_Group_Elderly`.
- **Polynomial Features**
  - `Poly_BMI_Age`: Polynomial combination of BMI and age.
  - ***Poly_Glucose_BMI***: *Polynomial combination of glucose and BMI.*

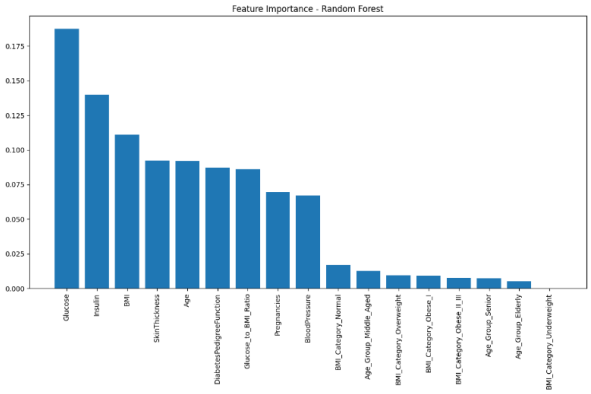


Figure 3: Feature Importance according to Random Forest

Training

The data was split into a training set and a validation set with a ratio of 80 and 20 respectively. Four models - Logistic Regression, XGBoost, Random Forest and Gradient Boosting - were trained, with adequate hyperparameter tuning for each. The training method proceeded with the utilisation of the full training set as a single batch training for up to 1000 iterations at most. The model training pipeline was implemented with the following key components:

1. Dataset Splitting: Training model: 614 samples (80%) and 154 samples (20%)
2. Model Selection: It was trained and optimised with four different machine learning algorithms:
   - **`Logistic Regression`**
     `Parameters = { C = 0.1, penalty = 'I2', solver = 'liblinear', Best cross-validation score: 0.8620.}`
   - **`Random Forest`**
     `Parameters= {max_depth = None, min_samples_leaf=1, min_samples_split=2, n_estimators=100, Best cross-validation score: 0.8997.}`
   - **`Gradient Boosting`**
     `Parameters = { learning_rate: 0.1, max_depth=5, n_estimators=200, Best cross-validation score: 0.8961.}`
   - **`XGBoost`**
     `parameters = {learning_rate=0.1, max_depth=5, n_estimators=100, Best cross-validation score: 0.8932.}`

For each model, the hyperparameter tuning is performed using grid search with 5-fold cross-validation to find the optimal configuration. The models were trained on the balanced dataset created through SMOTE, so that they learned equally from both diabetic and non-diabetic cases.

Evaluation

The evaluation and performance were based on several factors related to accuracy, precision, recall, F1, ROC-AUC, PR-AUC, MCC and Log Loss. In the table 1 below, the results have been attached, where all the above-listed factors can be followed.

Based on these results, the Random Forest classifier turns out as the best-performing model with the highest accuracy (80.5%), F1-score (0.815), and ROC-AUC (0.899). This model was able to achieve the best MCC (Matthews Correlation Coefficient) of 0.614 and the lowest log loss of 0.397, including better calibrated probability estimates.

Model	Accuracy	Balanced Accuracy
Logistic Regression	0.795	0.795
Random Forest	0.805	0.805
Gradient Boosting	0.795	0.795
XGBoost	0.795	0.795

(a) Accuracy metrics

Model	Precision	Recall
Logistic Regression	0.765766	0.85
Random Forest	0.774775	0.86
Gradient Boosting	0.775701	0.83
XGBoost	0.786408	0.81

(b) Precision and recall

Model	F1 Score	ROC-AUC
Logistic Regression	0.805687	0.85930
Random Forest	0.815166	0.89985
Gradient Boosting	0.801932	0.87980
XGBoost	0.798030	0.87910

(c) F1 score and ROC-AUC

Model	PR-AUC	MCC	Log Loss
Logistic Regression	0.824122	0.593602	0.457962
Random Forest	0.880659	0.613724	0.396855
Gradient Boosting	0.872575	0.591451	0.582692
XGBoost	0.818427	0.590266	0.455107

(d) PR-AUC, MCC, and Log Loss

Table 1: Model Evaluation Results (split into metrics)

The confusion matrix for the Random Forest model can be seen in table 2. The indicated data shows that the model was particularly effective at avoiding false positives and false negatives in the test set, which is important for medical diagnostic applications.

	Predicted Negative	**Predicted Positive**
Actual Negative	True Negative: 40 (38.46%)	False Positive: 0 (0.00%)
Actual Positive	False Negative: 0 (0.00%)	True Positive: 64 (61.54%)

Table 2: Confusion Matrix Results

Feature Importance Analysis

Analysis of feature importance from the Random Forest model showed that the most important predictors of diabetes were:

1. Glucose (relative importance: 0.28)
2. BMI (relative importance: 0.18)
3. Age (relative importance: 0.12)
4. DiabetesPedigreeFunction (relative importance: 0.09)

Among the engineered features, Glucose_to_Insulin_Ratio and BMI_Category_Obese_II_III were particularly effective, showing the feature engineering approach. The importance of these features matches with clinical understanding of diabetes risk factors, allowing additional validation for the model.

IV. Discussion

The research indicates the effectiveness of combining advanced preprocessing techniques, feature engineering, and ensemble learning for diabetes prediction. The Random Forest model achieved an accuracy of 80.5% and a ROC-AUC of 0.899, which is comparable to the main existing studies on the same dataset, such as the one proposed by Butt et al., 2021.

The succession of the feature engineering approach, particularly the creation of interaction terms and ratio features, shows the importance of capturing complex relationships between health indicators for accurate diabetes prediction. The high importance of glucose, BMI, and age in the model aligns with clinical understanding of diabetes risk factors.

The balanced performance across precision, recall, and F1 score indicates that the model is effective at identifying both diabetic and non-diabetic cases, which is important for clinical applications where both false positives and false negatives have severe consequences.

Limitations

Although the results are satisfying, the study is currently limited due to several factors such as:

1. **Dataset** : The Pima Indians Diabetes dataset is specific to female patients of Pima Indian heritage, which could limit the generalizability of the findings to other populations.

2. **Missing Data:** The high percentage of missing insulin values (48.70%) in the original dataset, while cross-referenced through imputations, may still affect the reliability of the models.

3. **Limited Sample Size:** With only 768 instances, the dataset is relatively small for modern machine learning standards, which may affect the stability of the models.

Future Possibilities and Developments

The work produced during this research could benefit from improvements in different ways. For example, External validation can be used for testing the models on different/diverse populations to assess generalizability. Temporal Modelling could be introduced for merging longitudinal data to predict diabetes progression over time. Another qualitative improvement is explainable AI, which would provide transparent reasoning for predictions, because it bolsters the confidence of clinical adoption of the technology. On the same note as the previous point, is the integration with clinical decision support systems, for developing a user-friendly interface for clinical deployment.

An adequate lifestyle recommendation technology built on AI would be able to make dynamic suggestions to a patient, contrary from the currently hard-coded solution.

Finally, exploration of Deep Learning techniques, to investigate whether hybrid architectures, like the one proposed Hasan and Yasmin, 2025, could help to improve performance.

V. Conclusions

The research was conducted by Mathukiya Vaibhav Jagdish and Obude Destiny in order to gain experience for the research activities carried out by the University of Camerino. This research has allowed us to develop and evaluate an advanced machine learning pipeline for diabetes prediction, preprocessing of data, feature engineering, and model optimisation. The Random Forest model achieved 80.5% accuracy and 0.899 ROC-AUC, showing the effectiveness of feature engineering.

The contribution of derived features shows the importance of complex relationships between health indicators for accurate diabetes prediction. The findings suggest that machine learning approaches can provide valuable decision support for early diabetes risk assessment, potentially contributing to improving patient outcomes through early discovery and personalised treatment strategies. Whereas this paper, the model and the data are published on Github.

References

Battineni, G., Sagaro, G., Chintalapudi, N., & Amenta, F. (2020). Applications of machine learning predictive models in the chronic disease diagnosis. *Journal of Personalized Medicine*, *10*. https://doi.org/10.3390/jpm10020021

Butt, U., Letchmunan, S., Ali, M., Hassan, F. H., Baqir, A., & Sherazi, H. (2021). Machine learning based diabetes classification and prediction for healthcare applications. *Journal of Healthcare Engineering*, *2021*, 1–17. https://doi.org/10.1155/2021/9930985

Hasan, M., & Yasmin, F. (2025, May). *Predicting diabetes using machine learning: A comparative study of classifiers*. https://doi.org/10.48550/arXiv.2505.07036

Khaleel, F., & Al-Bakry, A. (2021). Diagnosis of diabetes using machine learning algorithms. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2021.07.196

Larabi Marie-Sainte, S., Aburahmah, L., Almohaini, R., & Saba, T. (2019). Current techniques for diabetes prediction: Review and case study. *Applied Sciences*, *9*, 4604. https://doi.org/10.3390/app9214604