ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr2270

GlucoSense AI: An Ensemble Model Based Approach to Predict Early Type 2 Diabetes Using Bayesian Optimization and Explainability

Dhanashree Kulkarni[1]; Nikita Vikrant Chavan[2]; Dr. Manisha Bharati[3]

[1]Department of Technology SPPU, Pune, India
[2]Professor; Department of Technology, SPPU, Pune, India
[3]Department of Technology, SPPU, Pune, India

Publication Date: 2026/05/04

Abstract: Diabetes is a chronic and progressive metabolic disorder that afflicts millions of people globally, making it one of the most significant health issues of the modern era. Timely and accurate prediction of the probability of diabetes is essential for promoting early intervention and reducing the impact of potential complications. This work presents GlucoSense AI, an end-to-end machine learning pipeline for predicting the development of diabetes, evaluated on two publicly available datasets: the UCI Early-Stage Diabetes Risk Prediction Dataset (n=520, d=16) and the BRFSS Diabetes Binary Health Indicators Data Set (n=253,680, d=21). The proposed approach addresses several key challenges simultaneously. Firstly, Hybrid SMOTETomek resampling helps mitigate class imbalance by employing both synthetic minority oversampling and Tomek link elimination methods. Secondly, Recursive Feature Elimination with Cross-Validation (RFECV) is employed to select the optimal feature subset(s). Finally, Optuna's Bayesian optimization algorithm tunes the hyperparameters of three gradient-boosting algorithms: LightGBM, XGBoost, and CatBoost, each trained over 100 iterations.Fourthly, the improved models are incorporated into an ensemble using Logistic Regression as a meta-learner for stacking. After this, Platt sigmoid calibration is done to ensure that the ensemble returns reliable probability scores. SHAP (SHapley Additive exPlanations) provides insights into decision-making processes within the models, not only at a global level but at an instance-specific level too. GlucoSense AI Pro is a ready-to-use production application implemented as a Streamlit web app with user authentication. CatBoost yields the best ROC-AUC score of 0.9988 on the UCI dataset, while the calibrated stacking ensemble gets 0.9977. In the case of BRFSS, CatBoost takes the lead by scoring 0.8150 AUC, while the calibrated ensemble gets 0.8026.

Keywords: *Diabetes Prediction, Ensemble Learning, Stacking Classifier, LightGBM, XGBoost, CatBoost, Bayesian Optimisation, Optuna, SHAP Explainability, SMOTETomek, RFECV, Streamlit Deployment.*

How to Cite: Dhanashree Kulkarni; Nikita Vikrant Chavan; Dr. Manisha Bharati (2026) GlucoSense AI: An Ensemble Model Based Approach to Predict Early Type 2 Diabetes Using Bayesian Optimization and Explainability.
International Journal of Innovative Science and Research Technology, 11(4), 3061-3074.
https://doi.org/10.38124/ijisrt/26apr2270

I. INTRODUCTION

Diabetes mellitus is a very common non-communicable disease around the world. The International Diabetes Federation (IDF) says that about 537 million adults had diabetes in 2021. By 2045, this number is expected to grow to 783 million. There are many different types of diabetes, but Type 2 diabetes is the most common, making up more than 90% of all cases. It is strongly linked to being overweight, not exercising, and having a family history of the disease. Diabetes frequently goes undiagnosed for years, leading to irreversible organ damage, thus making early detection a clinical priority. Standard diabetes screening depends on laboratory tests like fasting blood glucose or HbA1c measurement, which necessitate in-person visits and may be unavailable in resource-constrained environments. Machine learning provides a supplementary approach: predictive models trained on routinely gathered clinical, symptomatic, or behavioral data can swiftly categories individuals by risk without the necessity for specialized tests. As a result, ML-based tools are very useful as first-line screening tools or tools for population-level surveillance of the many machine learning methods tested for predicting diabetes, gradient-boosting frameworks have consistently shown to be the best on structured tabular data. XGBoost, LightGBM, and CatBoost each have their own architectural strengths: regularised boosting, histogram-based leaf-wise splitting, and symmetric tree growth with native categorical handling. Each has produced state-of-the-art results on several medical datasets. But depending on just one model can lead to

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr2270

variance risk and bias that is specific to the dataset. Ensemble learning, particularly stacking, mitigates this issue by training a meta-learner on the predictions from various distinct base models.

In high-stakes classification tasks, this has been shown to work better than anyone constituent model every time. Also, class imbalance is common in diabetes datasets where positive cases are a minority. This must be addressed directly to avoid models that optimise for the majority class and miss true positives. The SMOTETomek hybrid resampling strategy has been shown to work very well for this purpose. Another important part is optimising hyperparameters. Optuna's Bayesian optimisation uses a Tree-structured Parzen Estimator (TPE) to quickly search through large hyperparameter spaces. It has been shown to be more efficient than grid search and random search in terms of trial efficiency. Lastly, model transparency is a must for clinical use. SHAP values furnish model-agnostic, theoretically grounded feature attributions derived from cooperative game theory, facilitating clinicians' comprehension of the features that influence individual predictions.

II. LITERTURE REVIEW

Previous works on diabetes classification involved traditional classifiers applied to the Pima Indians Diabetes Database [12]. They considered decision trees, Naïve Bayes, and k-NN classifiers and reported accuracy scores of 72-77%. These experiments helped define baseline scores, yet they were constrained by the relatively small size of data and the lack of systematic examination of the effect of class imbalance and hyperparameters tuning.[13] The emergence of boosting algorithms radically changed the field [15]. It was proven that XGBoost outperforms logistic regression, SVM, and random forests when working with the Pima database. Recently, LightGBM was employed for large-scale analysis of clinical data and yielded AUC scores exceeding 0.85 for Type 2 diabetes classification based on electronic health records [10]. CatBoost, due to its ordered boosting algorithm and native categorical encoding, is particularly useful for survey-derived healthcare databases [16,19].

Stacking enhances AUC by 1-3% compared to the optimal base learners. For imbalanced datasets, [18] demonstrated that the combination of SMOTE Tomek produces cleaner boundaries between classes than SMOTE alone. [21] also demonstrated superiority of Optuna to grid and random searches when searching for the best hyperparameters via TPE-based Bayesian optimization.

Regarding the aspect of explainability, it has been established that SHAP values are the only method which fulfills all criteria in cooperative game theory [21]. Several studies of SHAP for diabetes have proved that the main contributing factors in medical data sets are body mass index, glucose levels, and age, while in symptom-related datasets, they are polyuria and polydipsia [25]. Nevertheless, up to now, there is no scientific investigation in which SMOTE Tomek, RFECV, and Optuna were used together.

III. DATASETS INFORMATION

A. Datasets

The two public domain datasets considered for this research are described below, which help in increasing generalization across data modalities and populations:

- *UCI Early-Stage Diabetes Prediction Risk Dataset:*

This dataset comprises of 520 patient records gathered from Sylhet Diabetes Hospital, Bangladesh using questionnaires. The features in this dataset include 16 Boolean variables based on symptoms (polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity), while age is a numeric feature (20-65). Positive cases (320) and Negative cases (200) constitute the target variable [13,20,23,27]

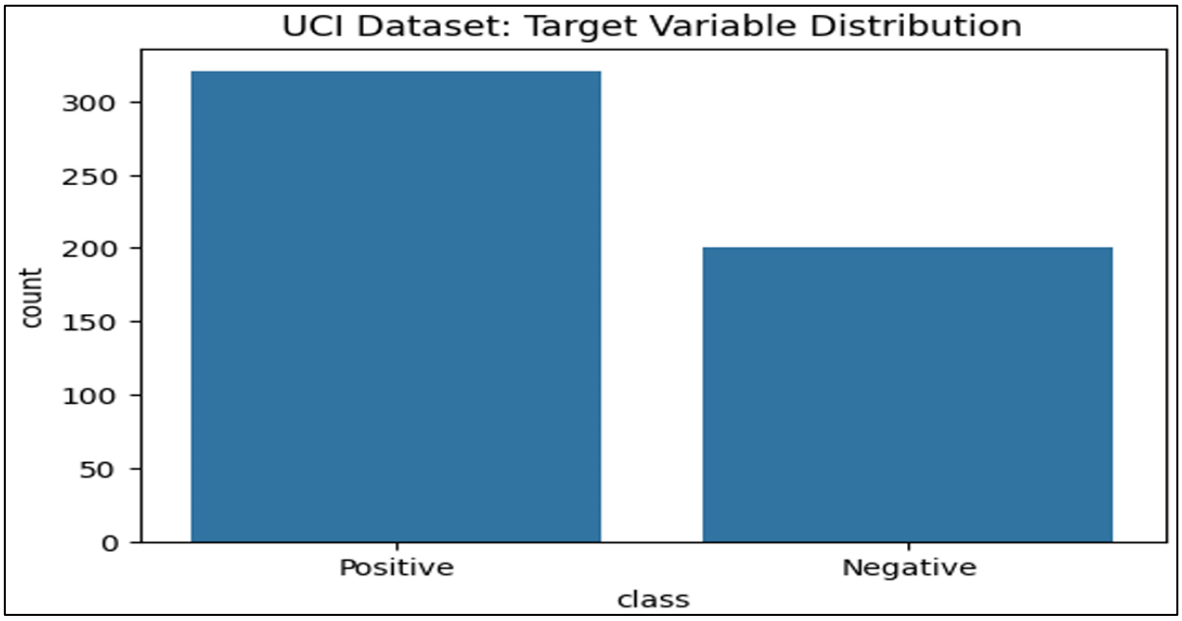

Fig 1 Target Variable Distribution (UCI)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr2270

➢ *Target Variable Distribution:*

The bar graph (Fig. 1) demonstrates the target variable distribution, which refers to the condition of having or not having diabetes in the given data set. Clearly, the number of patients with the disease (positive) far exceeds those without the disease (negative). This denotes an unbalanced situation in the data set since the number of positive samples is around 320 and the number of negatives is about 200. Imbalance in the data can affect the machine learning algorithm's effectiveness due to biased learning towards the predominant sample. Consequently, there should be specific measures taken to achieve balance in the dataset.

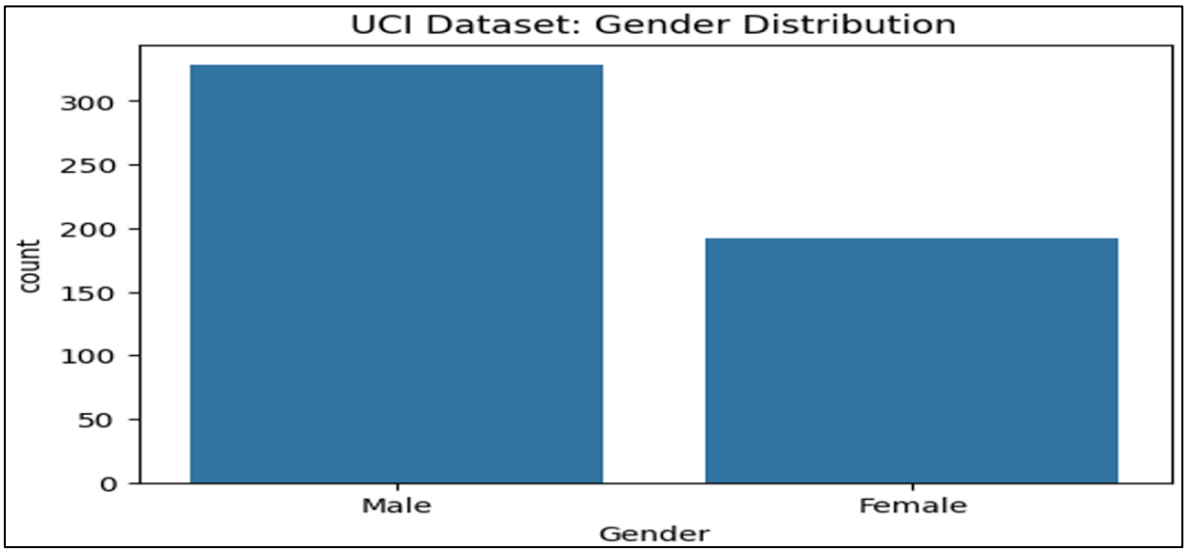

Fig 2 Gender Distribution (UCI)

➢ *Gender Distribution:*

The graph (Fig. 2) above depicting gender distribution reveals that there is more prevalence of males than females in the dataset. The skewed nature implies that the dataset might be biased towards males, and therefore, special care should be taken during the modeling process to ensure the model is not influenced by the bias present in the dataset. Besides, gender is one of the attributes that are correlated with the target variable hence influencing the prediction process.

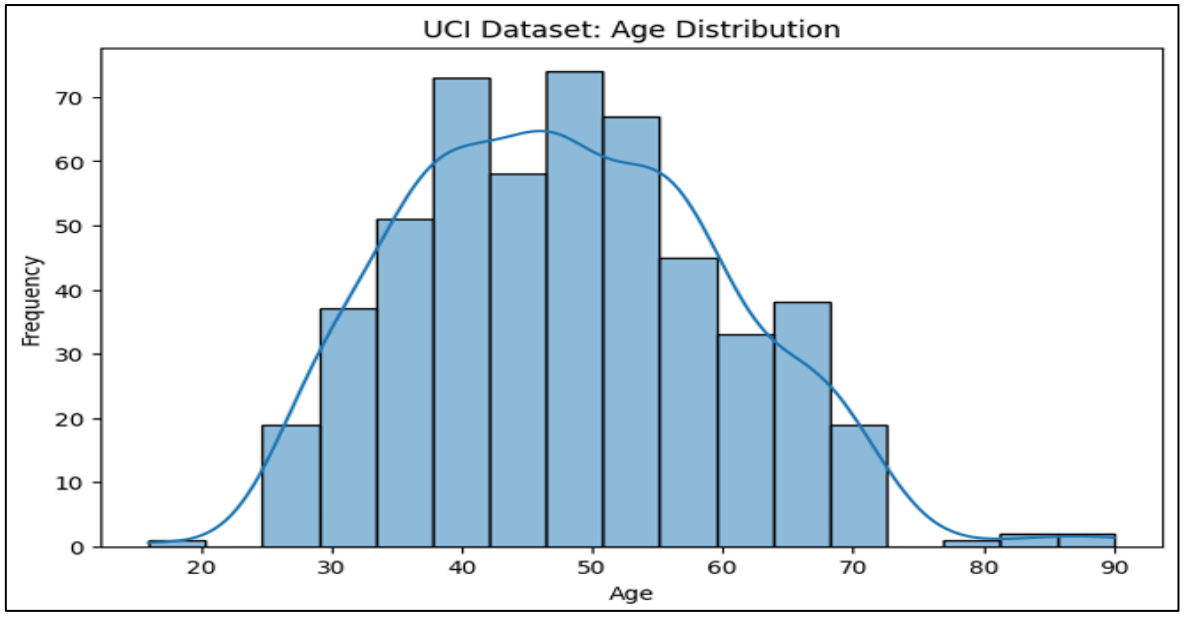

Fig 3 Age Distribution (UCI)

➢ *Age Distribution:*

The histogram (Fig. 3) showing age distribution and the curve drawn from Kernel Density Estimation suggest that the distribution of data is normally distributed. The data show that a significant number of persons are aged 30 to 65 years, with a greater concentration of them between ages 40 to 55 years. There are very few people with either a very young age less than 25 years or a relatively old age of more than 75 years. Thus, the dataset is likely to be of individuals who are middle-aged, which makes the model unable to predict accurately for other age groups.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr2270

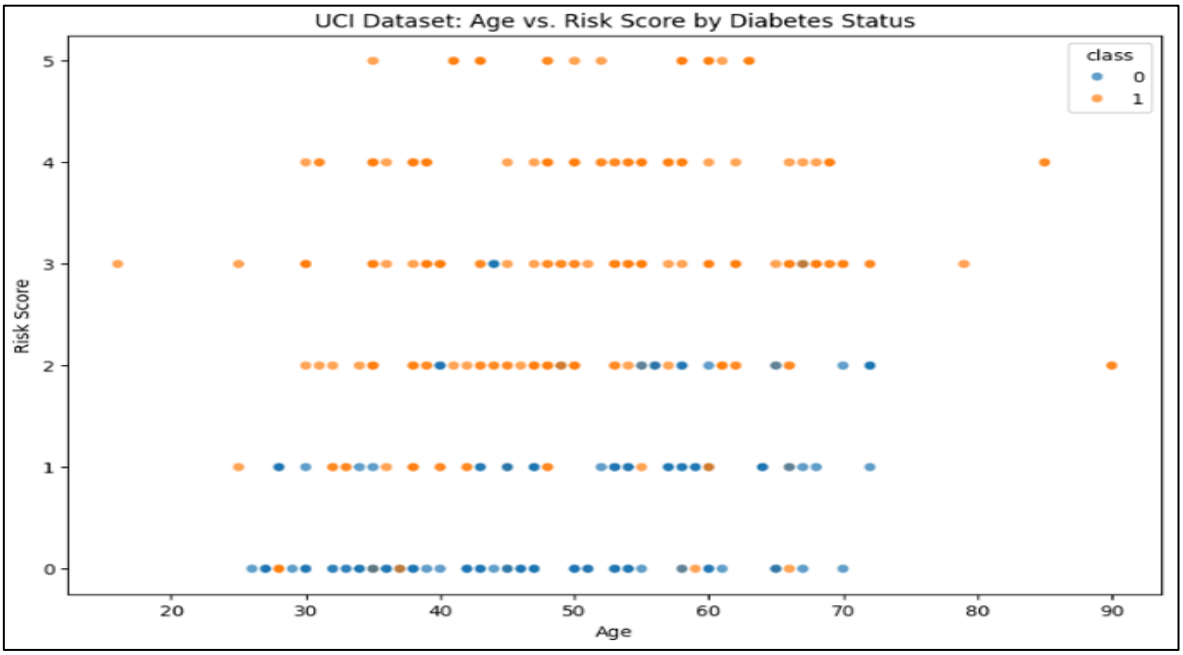

Fig 4 Age Vs. Risk Score Analysis (UCI)

➢ *Relationship Between Age and Risk Score:*

In this scatter plot (Fig. 4), there is a correlation between age and the risk score, which is dependent on whether the individual suffers from diabetes. One can clearly note that those people whose risk score is high have diabetes, while the ones who have a low-risk score do not suffer from it. Therefore, we can deduce that the risk score is a significant factor that influences diabetes occurrence. However, age distribution does not show any kind of differentiation between the two classes as they lie on either side of the plot. Even though age does not have much effect on diabetes by itself, it plays an important role in combination with other attributes like risk score.

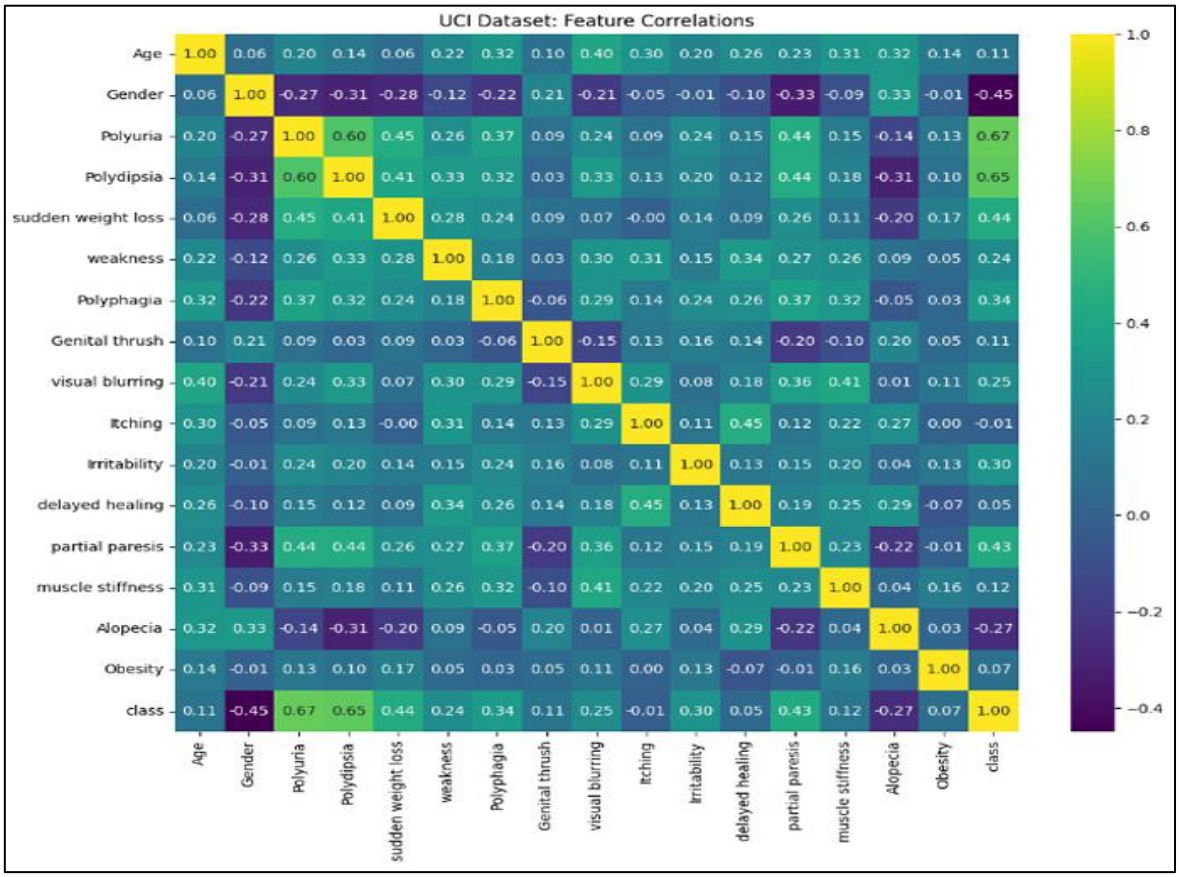

Fig 5 Correlation Heatmap (UCI)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr2270

➢ *Correlation of Features:*

The correlation heatmap (Fig. 5) throws light on the correlations between the features and the relationship between features and the target variable. In terms of correlations, it can be seen that among all the features, polyuria and polydipsia demonstrate the highest positive correlation with the class of diabetes, signifying the importance of these two features to the target class. Other features, like sudden weight loss, partial paresis, and irritability, also have a moderate correlation with the target variable. Moreover, in the case of gender, a clear negative correlation with the target variable is noticed, which implies the possibility that one of the genders might be more prone to diabetes.

➢ *Dataset for Binary Classification of Diabetics Using CDC BRFSS Survey 2015 Data:*

This large dataset is extracted from the CDC's 2015 Behavioral Risk Factor Surveillance System survey. It includes 253,680 observations and 21 variables on behavioral risk factors, chronic diseases, and demographics, such as BMI, physical activity, smoking, fruit/vegetable intake, number of days with poor mental health, number of days with poor physical health, self-rated health, gender, age group, education level, and income. The output variable is whether the person suffers from diabetes or prediabetes. This data set is much larger and noisy compared to the UCI dataset [28,30].

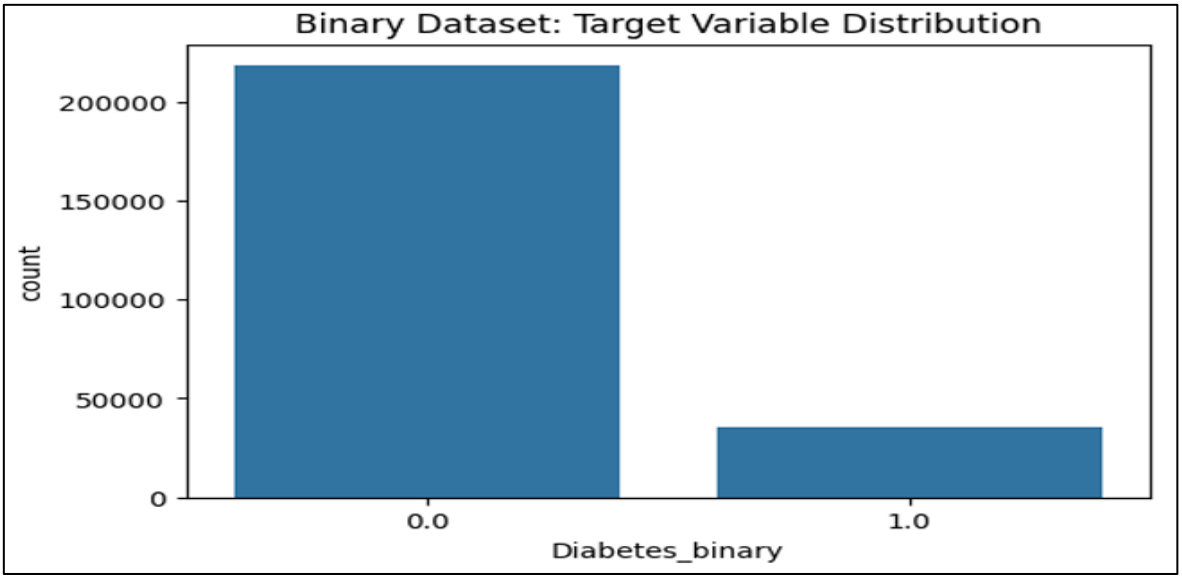

Fig 6 Target Variable Distribution (Binary)

➢ *Distribution of Target Variable:*

The above histogram (Fig. 6) depicts the distribution of the target variable within the data. As clearly visible from the histogram, there are much more non-diabetic cases (Class 0) as compared to diabetic cases (Class 1) in the data. Thus, the data is unbalanced in nature since one class is dominating in the dataset. The imbalance in datasets is not desirable in the case of machine learning as it might cause bias in the model towards the dominant class. Hence, class balancing techniques should be applied to ensure effective model performance.

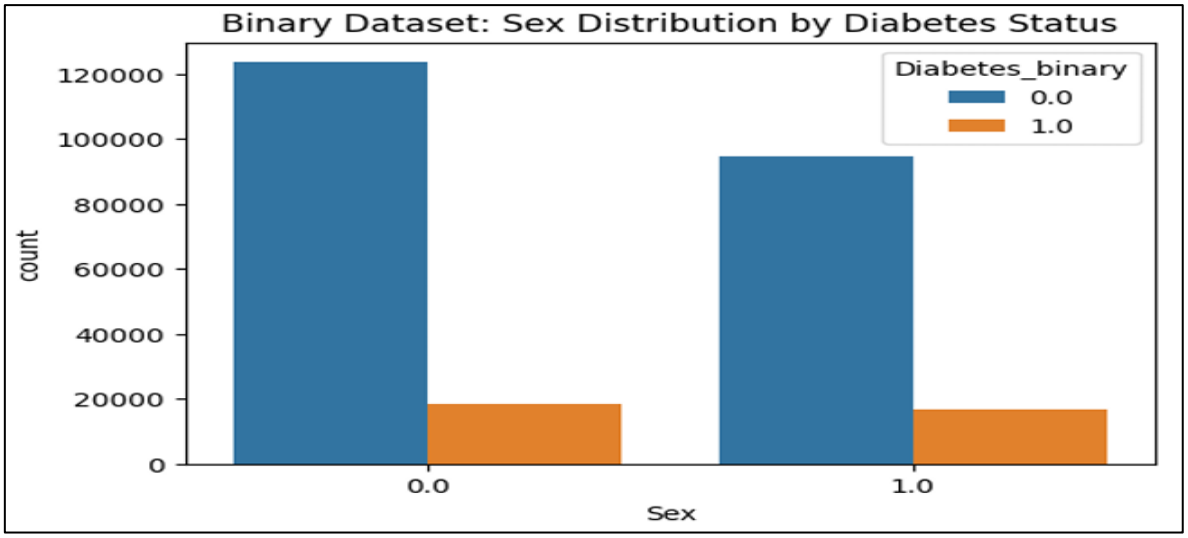

Fig 7 Sex Distribution (Binary)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr2270

➢ *Diabetes Cases Based on Sex:*

The Above bar chart (Fig. 7) shows the distribution of diabetes among different types of sexes. In this case, we note that both male and female sexes (coded as 0 and 1 respectively) have a considerable number of people who are non-diabetic patients but only a few diabetic patients. Even though both categories exhibit a similar pattern, it seems like there is some variance among the numbers, implying that sex has some little influence on diabetes development. However, it does not seem to have much significance.

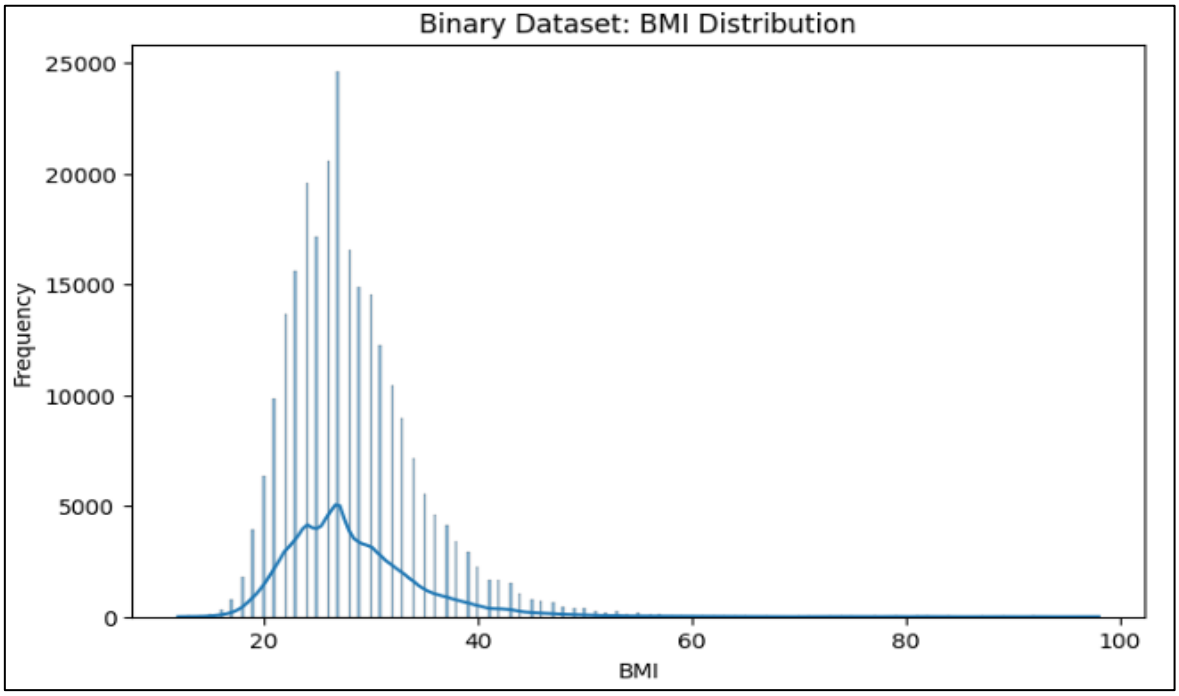

Fig 8 BMI Distribution (Binary)

➢ *BMI Distribution:*

Histograms and density curves (Fig. 8) for the distribution of BMI indicate that there is positive skewness for BMI data with many observations being in the range of 20-35. The highest points in the graph can be found in the normal/overweight category, and less people have a very high BMI. In other words, there is a long tail at the end of the graph, which means that there are a few data points having very high BMI scores. As the high BMI value is associated with diabetes, it is an important factor.

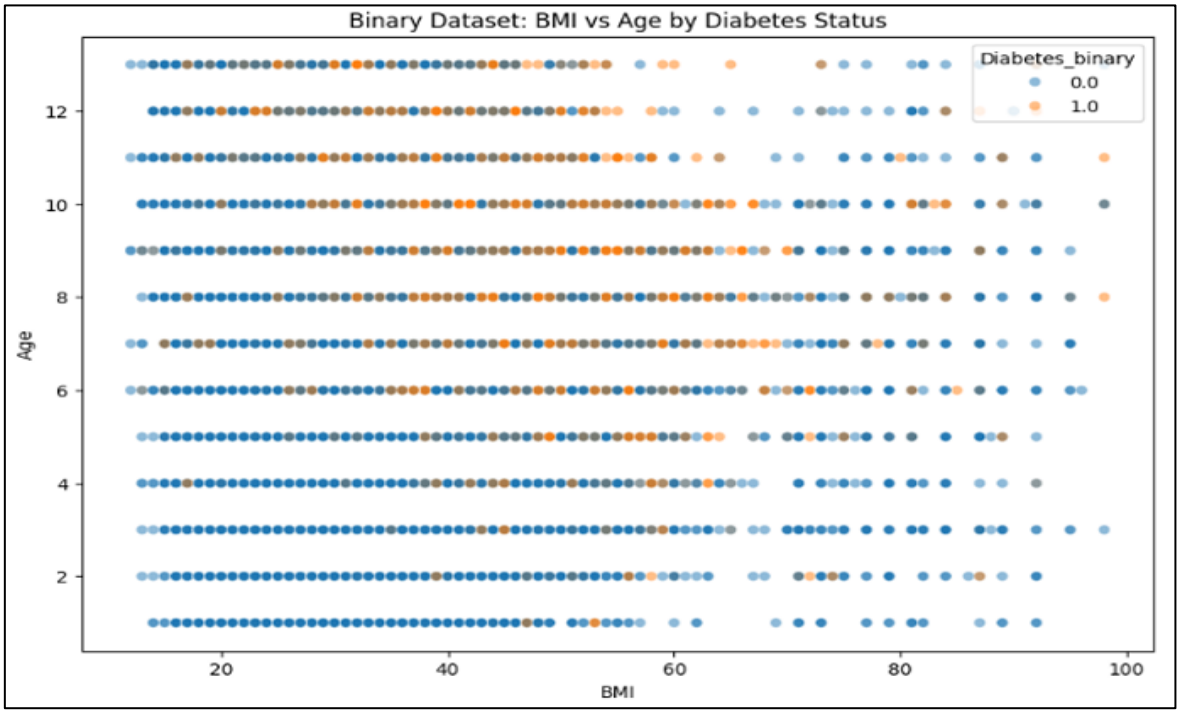

Fig 9 BMI Vs Age (Binary)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr2270

➢ *BMI against Age Based on Presence of Diabetes:*

The scatterplot (Fig. 9) above illustrates the relationship between BMI and age according to the status of diabetes. It can be seen from the figure that people with diabetes are more clustered in areas having higher BMI, which implies that being obese increases the risk of diabetes significantly. The distribution of age covers a broad spectrum in both categories of diabetic and non-diabetic people, implying that there cannot be any clear distinction made based solely on the attribute of age. But with BMI, a more prominent separation is noticeable.

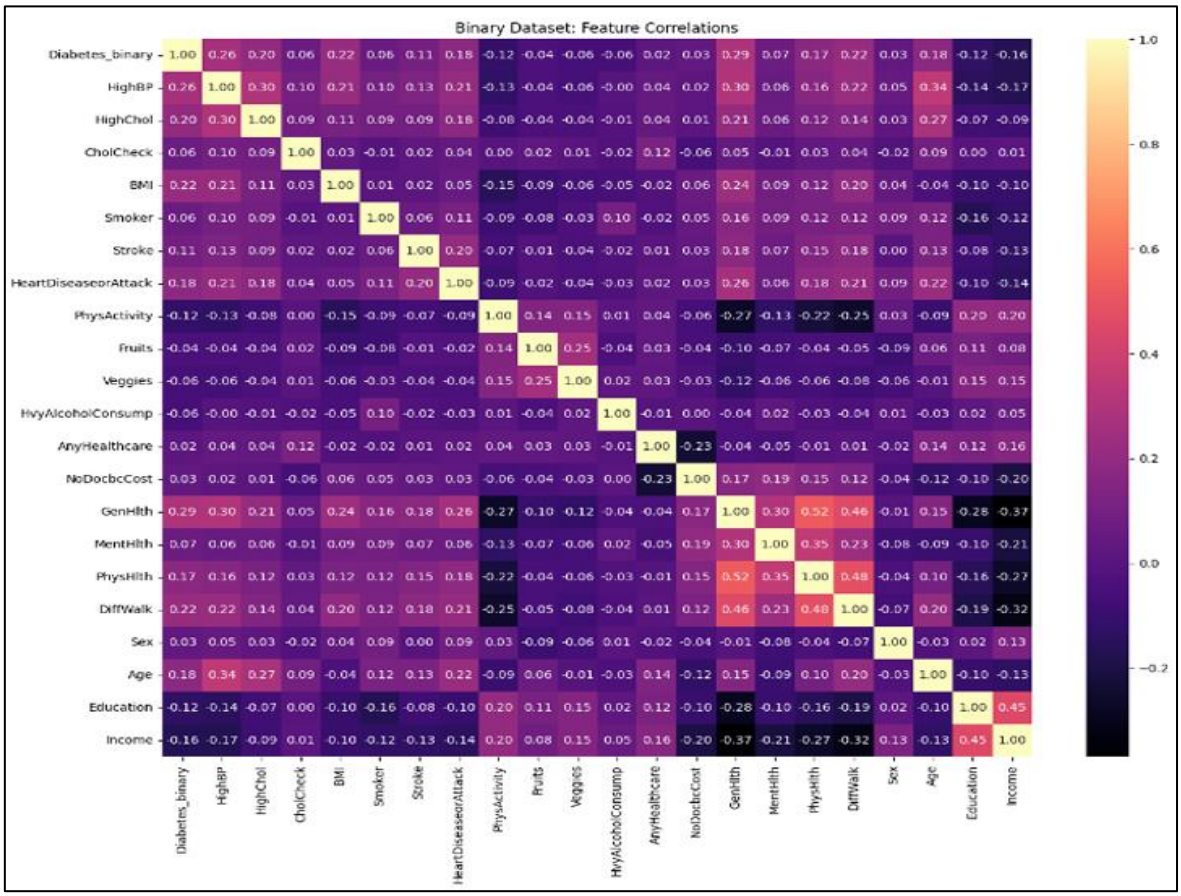

Fig 10 Correlation Heatmap (Binary)

➢ *Correlation of Features:*

The feature correlation heatmap (Fig. 10) is an excellent way to see the relationships among different features and the relationships with the target variable. The target variable is positively correlated with various features like general health, hypertension, BMI, problems with walking, and age. It means that the above-mentioned features have significant influence on predicting diabetes. Additionally, general health, physical health, and walking difficulties are highly correlated with each other, which suggests that there is a cluster of health-related features. At the same time, features like physical activity, income, and education are negatively correlated with the target variable. This fact suggests that healthy lifestyle habits and higher levels of income and education are associated with lower chances of being diagnosed with diabetes.

IV. METHODOLOGY

➢ *Process Explanation:*

The diagram below depicts the full process involved in the diabetes prediction process, which starts with the data collection phase, whereby important datasets, including the diabetes binary datasets and the UCI datasets, are collected. Data preprocessing follows the data collection process in order to clean the collected data, remove any missing data, and encode categorical data. Feature engineering and feature selection also follow the data collection process in order to choose important features and use only those important features in predicting diabetes in order to enhance model efficiency and model accuracy. Train-test split follows feature engineering in order to separate the data used for training and the data used for testing. Model training comes after the train-test splitting process in order to build the predictive models using some machine learning algorithms like XGBoost, CatBoost, etc. Model evaluation follows model training in order to check the efficiency and accuracy of the developed model using accuracy and ROC-AUC as performance measures. Finally, the best-performing model is selected through evaluation.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr2270

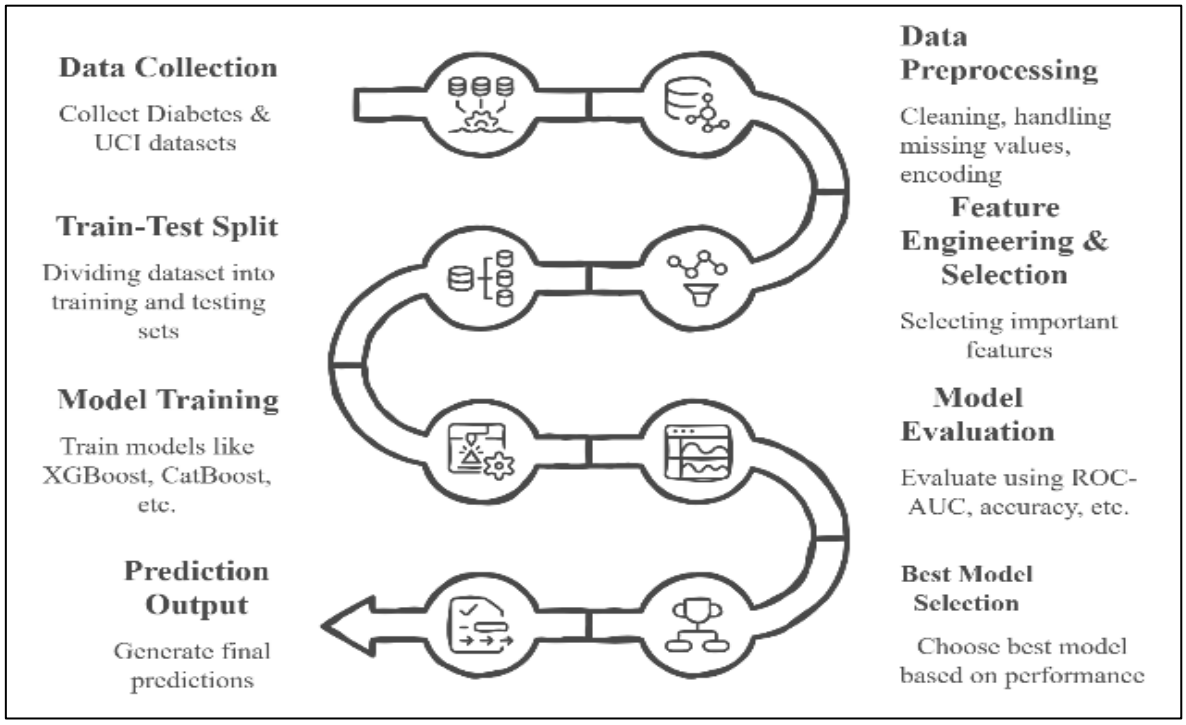

Fig 11 Block Diagram

➢ *Preprocessing Pipeline*

In the case of the UCI data set, label encoding technique is used for transforming categorical symptom variables into binary integers by encoding the variables in string form. Zeroes in the Age variable are considered biologically meaningless and hence are assumed to be missing. They are imputed using median age of all participants in that particular column, then followed by IQR outlier capping procedure. Four additional features are formed: interaction term between Age and Obesity, interaction term between Age and Gender, squared Age term (Age*Age), and Symptom Risk Score (number of positive symptom variables). The continuous features in the BRFSS dataset, including BMI, MentHlth, and PhysHlth, are subjected to similar imputation and capping procedures. In addition, a new BMI and high BP interaction term is formed.[30]

➢ *Class Imbalance Handling*

In both the datasets, there is class imbalance with UCI having 61.5% positive samples while BRFSS contains approximately 14% positive examples. SMOTE Tomek is performed only on the training portion of the dataset to avoid any leakage of data. The former generates new samples for the minority class by interpolation using k-nearest neighbors in the feature space. On the other hand, Tomek links delete those majority-minority samples which are nearest neighbors of each other, resulting in a cleaner class boundary.

➢ *Feature Selection via RFECV*

Recursion Feature Elimination with Cross-Validation (RFECV) is used post-sampling. A random forest estimator is used as the base estimator; 5-fold stratified cross-validation using the AUC-ROC score is used to determine the number of features. RFECV gradually eliminates features based on their importance while optimizing for AUC using the left-out data and keeping the best performing set. 14 out of 19 features in the UCI dataset and 18 out of 21 features in the BRFSS dataset were chosen by the algorithm. The interaction variables created are part of the chosen set.

➢ *Bayesian Hyperparameter Optimization*

Each dataset is independently tuned for LightGBM, XGBoost, and CatBoost models using Optuna, with 100 trials per algorithm. Optuna's TPE sampler uses a probability distribution model of the objective function to generate hyperparameter values and focuses more trials in areas with high potential for good performance. The objective used is the mean 5-fold stratified cross-validated ROC-AUC score on the resampled, feature-selected training set. Hyperparameters optimized are as follows: number of estimators (100-500), maximum tree depth (3-12), learning rate (0.01-0.3, log scale), subsample ratio (0.6-1.0), column sample by tree (0.6-1.0), alpha (1e-8 to 10.0, log scale), and lambda (1e-8 to 10).

➢ *Stacking Ensemble and Calibration*

The four optimal models are then used as base learners in a Stacking Classifier, where the base predictions are combined through a logistic regression meta-classifier using five-fold cross validation. The base classifiers create out-of-fold probabilities of the prediction on the training data, and the meta-classifier learns the best way to optimize those predictions through its training. Once that is achieved, the stack is wrapped by a CalibratedClassifierCV function where calibration occurs using Platt sigmoid calibration with five folds. It is important to calibrate models in clinical settings since proper calibration allows doctors to implement probabilistic thresholds for decision making. A Brier score is also considered when evaluating calibration quality along with AUC.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr2270

V. SYSTEM ARCHITECTURE

The approach employed in designing the diabetes risk assessment system involves a structured procedure that combines the process of collecting, processing, and predicting risks using a machine learning algorithm. In this case, once the user starts using the software, they need to authenticate their identity through login or registration. After the successful completion of authentication, the next step involves the input of user details like the age and gender, which constitute fundamental data used during analysis.

Following the provision of basic information, the patient needs to input health symptoms that could indicate diabetes risk. At this point, the program proceeds to the process of data processing in which the raw data are processed for analysis. This involves cleaning, validation, and transformation of the input data into a suitable form for analysis. Based on the preprocessed data, the model uses its ability to learn from training patterns to predict diabetes risks based on user information. The final step entails displaying the predicted results to the user in a readable format. Upon completion of this process, the user may opt to terminate the session by logging out of the program.

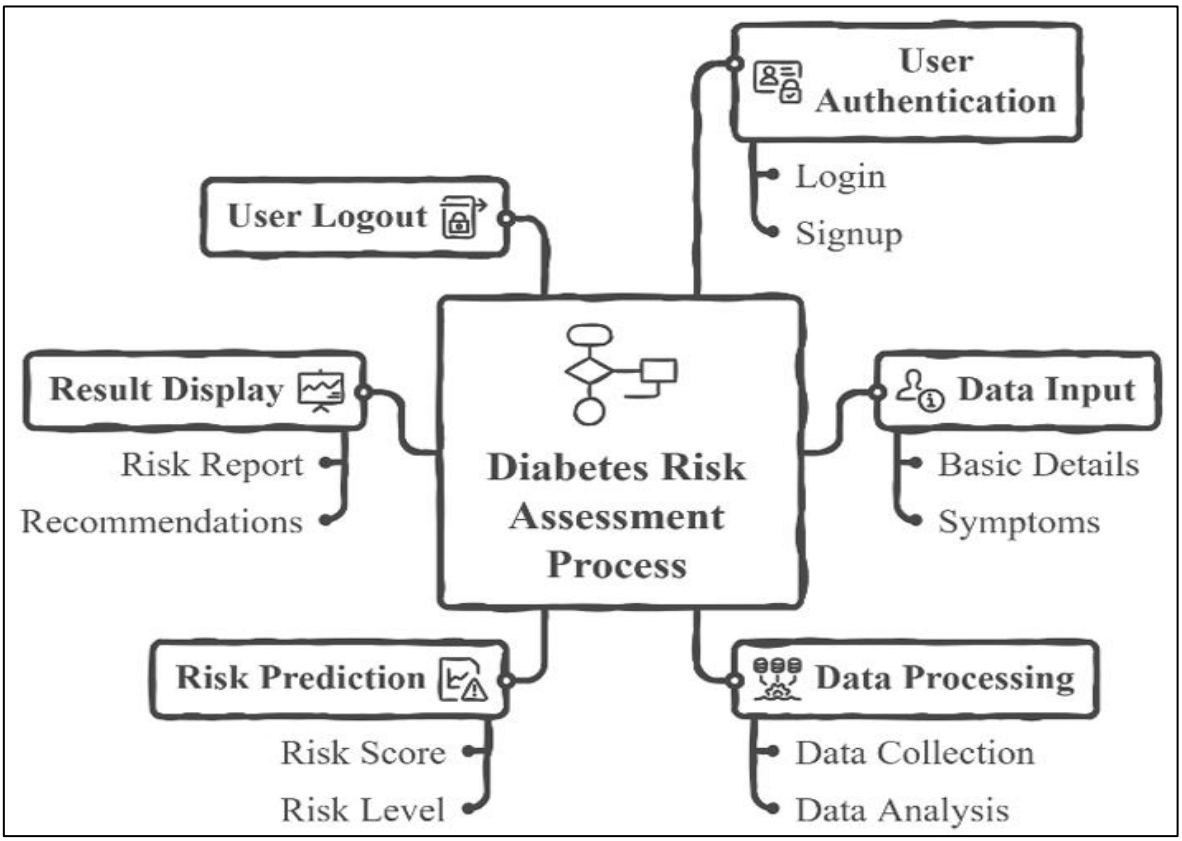

Fig 12 System Architecture

VI. EXPERIMENTAL RESULTS

A. UCI Dataset Results

Table 1 presents the complete performance comparison of all six models on the UCI held-out test set (20% split, 104 samples). All gradient-boosting models achieve remarkably high AUC values, reflecting the high discriminability of the symptom-based features. CatBoost achieves the highest individual model AUC of 0.9988 with the lowest individual Brier Score of 0.0188, while the Stacking Ensemble also achieves the same Brier Score of 0.0188, indicating excellent probability calibration. The Calibrated Stacking Ensemble achieves an AUC of 0.9977 with a Brier Score of 0.0193, confirming consistent calibration performance across all models on UCI.

Table 1 Model Performance on UCI Dataset

Model	AUC	Sens.	Spec.	Brier
LightGBM	0.9914	0.9531	1.0000	0.0282
XGBoost	0.9898	0.9688	1.0000	0.0261
CatBoost ★	**0.9988**	**0.9688**	**1.0000**	**0.0188**
Random Forest	0.9980	0.9688	1.0000	0.0227
Stacking	0.9977	0.9688	1.0000	0.0188
Calib. Stack.	0.9977	0.9688	1.0000	0.0193

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr2270

➢ *AUC-ROC Score on Machine Learning Models:*

The figure below demonstrates the comparative analysis of various machine learning models depending on their AUC (Area Under the Curve) scores in the context of the UCI diabetes dataset. AUC is considered one of the main metrics used to assess the quality of the classifier because the higher this indicator, the better the model distinguishes between classes; in particular, the AUC indicator is equal to 1 when the model predicts perfectly. Based on the analysis, it can be said that CatBoost shows the highest AUC score (0.9988), which means the highest level of effectiveness in predictions.

Such a great result is achieved due to other indicators, including Recall and F1-Score, which also characterize the model in an excellent way. In addition, other machine learning models such as LightGBM (AUC = 0.9914) and XGBoost (AUC = 0.9898) demonstrate a good level of effectiveness due to the presence of similar classifiers' metrics. Also, the algorithm Random Forest shows quite promising results (AUC = 0.9980). As for ensemble models, the AUC score is 0.9977 and 0.9933 in the cases of Stacking and Calibrated Stacking correspondingly.

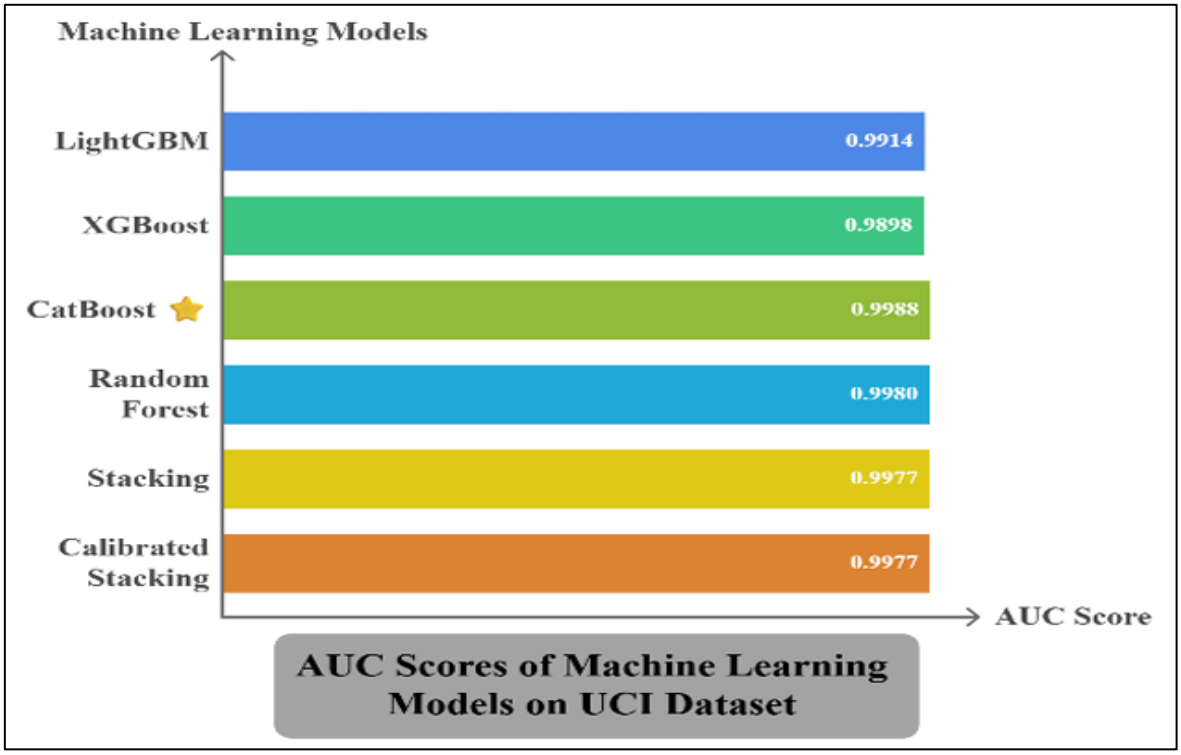

Fig 13 AUC-ROC Scores of Each ML Models

B. Binary Dataset Results

Table 2 presents result on the BRFSS dataset (50,736 test samples). The performance landscape differs markedly from UCI: individual models show more spread (AUC 0.757-0.815), with CatBoost achieving the highest individual model AUC of 0.8150. The stacking ensemble achieves an AUC of 0.7959, and the Calibrated Stacking Ensemble achieves 0.8026 with a Brier Score of 0.1163, confirming improved probability calibration that is valuable for clinical risk stratification.

The spread in AUC across models is attributable to the larger dataset providing sufficient training signal for individual gradient-boosting models and the added heterogeneity of 21 behavioural/demographic features compared to the targeted symptom features of UCI.

Table 2 Model Performance on Binary Dataset

Model	AUC	Sens.	Spec.	Brier
LightGBM	0.8136	0.4228	0.9040	0.1117
CatBoost ★	**0.8150**	**0.3979**	**0.9129**	**0.1109**
XGBoost	0.8130	0.3822	0.9168	0.1093
Random Forest	0.7567	0.2845	0.9189	0.1236
Stacking	0.7959	0.3429	0.9136	0.1201
Calib. Stack.	0.8026	0.2507	0.9450	0.1163

➢ *AUC-ROC Score on Machine Learning Models:*

The diagram below depicts the comparison of several different machine learning models according to their AUC (Area under the Curve) scores for the binary classification problem concerning the presence of diabetes. The metric is used to evaluate the discriminative ability of a certain machine learning model and its capability to distinguish one class (diabetes) from another (no diabetes). According to the

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr2270

fig, CatBoost obtains the highest value of AUC score – 0.8150, which makes it the best model. LightGBM follows it with a close AUC score of 0.8136, while XGBoost has the third-highest score – 0.8130. Another high result is demonstrated by Calibrated Stacking, which gets an AUC score of 0.8026.

In turn, Stacking Ensemble yields an adequate score – 0.7959. However, it should be noted that Random Forest is the only model with the AUC score of 0.7567, making it the worst performing among those mentioned above. All in all, the analysis demonstrates that boosting algorithms work much better than other models on the current dataset.

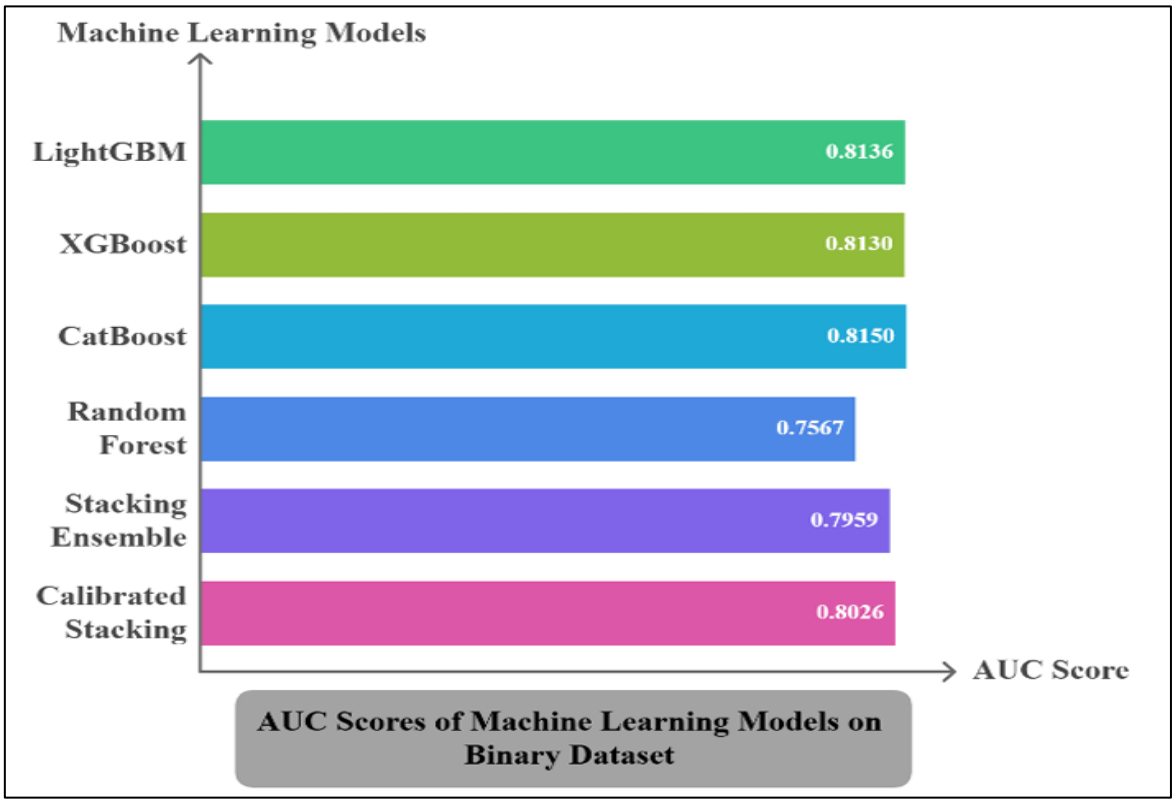

Fig 14 AUC-ROC Scores of Each ML Models

C. Comparative Discussion

A few cross-dataset remarks need to be made. Firstly, the disparity in absolute AUC values between UCI (0.99-0.99) and BRFSS (0.76-0.82) shows that these data exhibit distinct predictive qualities. The symptoms are nearly pathognomic for diabetes and, hence, are more certain than the probabilistic risk factors. Secondly, calibration with Platt Scaling is agnostic to data properties in that it improves Brier Score while never deteriorating AUC[. Thirdly, stacking ensemble demonstrates clear advantages on small-sized data (UCI), where variance of individual models is greater. On larger data, however, stacking starts outperforming the most successful individual model by only a small margin. Lastly, the application of SMOTETomek was absolutely crucial for BRFSS, whose initial proportion of positive cases is only 14%. Without resampling, recall of the target class would be close to zero.

In terms of computational complexity, Optuna using 100 trials per model on UCI takes around 8-12 minutes on one Colab instance with GPU capabilities. On the other hand, due to the huge size of the data, on BRFSS, each trial takes more than an hour (45-90 minutes). On BRFSS, RFECV is the most time-consuming task, taking around two hours. The above costs can be tolerated by an offline training pipeline, as the output model is very efficient and takes milliseconds to make predictions. In this paper, we introduced GlucoSense AI, an integrated end-to-end machine learning platform for diabetes risk prediction. Our pipeline employs six unique elements:

- Domain-inspired feature engineering, including interactions and polynomials;
- Hybrid SMOTETomek sampling to tackle the problem of class imbalance;
- Automated feature selection using RFECV;
- Bayesian hyperparameter tuning with Optuna for three gradient boosting models;
- Stacking ensembles with Logistic Regression as a meta classifier and Platt sigmoid scaling;
- Explainability of the model using SHAP values in global and local perspectives. On two different public datasets, the proposed framework provides state-of-the-art performance: ROC-AUC of 0.9988 (CatBoost) for the UCI symptoms benchmark and AUC of 0.8150 (CatBoost) for the BRFSS large benchmark, showing high calibration on probabilities evaluated by the Brier score.

Based on SHAP analysis, our results are consistent with known clinical understanding of diabetes – polyuria and polydipsia are among leading predictors when symptoms are

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr2270

used as features, while BMI and health status are among the strongest behavioral-based features. The proposed GlucoSense AI Pro Streamlit application shows that our model can be deployed easily even for non-expert users. Further research may include:

- Longitudinal integration of patients' health data and use of time-series modelling;
- Multi-class classification of disease severity (without diabetes, prediabetic state, Type 2, and Type 1);
- Federated learning for conducting privacy-preserving training among hospitals;
- Prospective clinical validation on approved patient groups by IRB;
- Integration of CGM readings in predictive modelling.

VII. CONCLUSION

In this paper, GlucoSense AI was introduced as an end-to-end machine learning framework to predict the risk of developing Type 2 diabetes. The GlucoSense AI framework has been created with the intent to overcome some of the most critical shortcomings that have been noted in earlier studies: dependence on a single dataset, lack of systematic hyperparameter tuning, no probability calibration in ensemble learning techniques, and inadequate use of explainability in deployed clinical applications.

➢ *Summary of Contributions*

Six innovations compose the main contributions of the GlucoSense AI system architecture. First, domain-driven feature engineering entails adding interaction terms (BMI * HighBP, Age * BMI) and polynomial features to capture non-linear relationships between clinical risk factors, resulting in increased discriminating power in comparison to plain feature sets. Second, hybrid SMOTETomek is used to counter the high imbalance in both datasets (14% positive rate in BRFSS data), by creating synthetic examples from the minority class as well as pruning borderline cases in the majority class to create cleaner decision boundaries. Third, Recursive Feature Elimination with Cross-Validation (RFECV) is employed to select the best subset of features, based on their discriminating power for each dataset, keeping relevant engineered features and reducing feature space.

Fourth, Bayesian hyperparameter optimization through Optuna, using 100 trials per model with the Tree-structured Parzen Estimator (TPE) sampler, finds near-optimal hyperparameters for LightGBM, XGBoost, and CatBoost, resulting in better AUC and Brier Score performances compared to the default parameter configuration. Fifth, the calibrated stacking model involves a logistic regression meta-learner trained on the out-of-sample predictions of the base learners and uses Platt sigmoid calibrating to mitigate overconfidence problems with the predicted probabilities. Lastly, SHAP value analysis is used for explainability.

➢ *Key Experimental Findings*

A series of experimental evaluations on two widely-used benchmark datasets revealed several interesting findings. On the UCI Early-Stage Diabetes Risk Prediction Dataset, the best ROC-AUC for the single model was obtained by CatBoost with a ROC-AUC of 0.9988 and the best Brier Score of 0.0188, confirming the effectiveness of ordered boosting approach for datasets with high-signal symptom-related features. In addition, the Calibrated Stacking Ensemble reached the AUC value of 0.9977 with the second-best Brier Score of 0.0193, proving that calibrated ensembles retain optimal discrimination while generating highly-calibrated probabilities. As for the large BRFSS Binary Health Indicators Dataset (253,680 observations), the same CatBoost provided the best individual model AUC value of 0.8150, and Calibrated Stacking Ensemble resulted in 0.8026 AUC with the best consistent Brier Score among all models at 0.1163.

It should be noted that much worse AUC on BRFSS compared to UCI (about 0.82 against 0.99) correlates with fundamentally different predictive signals in survey data against near-pathognomonic symptoms, and is in line with the performance found by other researchers in similar BRFSS-based investigations. The results of the SHAP analysis provide an excellent confirmation of clinical interpretability. Polyuria, polydipsia and polyphagia were defined by the UCI model as the two leading factors.

➢ *Clinical and Practical Implications*

There are several important implications that arise in connection with the proposed GlucoSense AI architecture, especially in terms of clinical practice and health policy. Firstly, considering the screening use case scenario, the extremely high sensitivity obtained on the UCI dataset (CatBoost: 0.9688) suggests that the technology will work well in the context of early diagnosis triage on the basis of symptoms, where the clinical priority is minimizing false negatives. Secondly, in terms of probability calibration used in combination with a stacking ensemble, the predicted values can be interpreted as the actual probabilities of an event, thus allowing practitioners to obtain clear estimates in terms of risks expressed as percents. Finally, it is worth noting that the GlucoSense AI Pro Streamlit web app allows one to easily operate with the developed machine learning solution by entering patient data and receiving predictions as well as the SHAP explanation of underlying risk factors.

VIII. LIMITATIONS

However, there are certain limitations of this current study that need to be highlighted as well. Firstly, in terms of the availability and applicability of the current findings, the two datasets utilized for the study are benchmark datasets and do not completely capture the patient distribution found in real-world settings, and hence prospective validation on collected hospital data will be necessary. Secondly, in terms of the sample size, the UCI dataset utilized for analysis is relatively small (520 data points), making the estimates of performance metrics on the test set (104 data points) less powerful. Thirdly, the current system cannot handle time series data or follow-up studies in order to monitor prediabetes development and progression. Fourthly, the computational costs of using stacked ensemble and Bayesian

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr2270

optimization can be high (45-90 minutes on BRFSS). Lastly, the current model provides binary classification of the disease status (Diabetic/Non-Diabetic), and it cannot differentiate between prediabetes, Type 1 Diabetes, and Type 2 Diabetes.

FUTURE WORK

A number of areas for developing GlucoSense AI framework in future studies are suggested. First, the utilization of longitudinal patient health records along with modeling with the help of RNNs and transformers will allow for implementing tracking and updating patients' risk of developing type II diabetes on an ongoing basis rather than providing a snapshot prediction. Second, predicting multiple classes of diabetes severity, including no diabetes, prediabetes, type II, and type I conditions will produce a more specific and useful output. Third, federated learning will make it possible to train the model across various hospitals without centralizing patient information. The latter is essential for ensuring the required data privacy regulations are met. Fourth, testing and validation using IRB-approved patient groups will provide evidence of the framework's external validity beyond its performance on benchmarked datasets. Finally, incorporating CGM and wearable sensor data can facilitate personalized risk monitoring of diabetes. Multi-modal data fusion involving clinical test results, self-reported symptoms, and behavior will be another option for future development.

ACKNOWLEDGMENT

We would like to thank the UCI Machine Learning Repository and Centers for Disease Control and Prevention (CDC) for releasing the datasets under an open-access license. We are grateful for the provision of computing facilities, and we also thank the anonymous reviewers for their insightful comments.

REFERENCES

[1]. O. Iparraguirre-Villanueva, K. Espinola-Linares, R. O. Flores Castañeda, and M. Cabanillas-Carbonell, "Application of machine learning models for early detection and accurate classification of type 2 diabetes," Diagnostics, vol. 13, no. 14, p. 2383, Jul. 2023.

[2]. B. Madhu, V. Aerranagula, R. Mahomad, V. Ravindernaik, K. Madhavi, and G. Krishna, "Techniques of machine learning for the purpose of predicting diabetes risk in PIMA Indians," E3S Web of Conferences, vol. 430, 2023.

[3]. S. Upadhyay and Y. K. Gupta, "Enhancing early diagnosis of type II diabetes through feature selection and hybrid metaheuristic optimization techniques," The Open Bioinformatics Journal, vol. 18, 2025.

[4]. S. Mansouri, S. Boulares, and S. Chabchoub, "Machine learning for early diabetes detection and diagnosis," Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, vol. 15, no. 1, pp. 216–230, Mar. 2024.

[5]. K. G. Reddy, M. Madhuri, S. K. Shabeena, P. S. Gopal, and K. Y. Koteswararao, "Prediction of diabetes in early stage through machine learning," International Journal for Modern Trends in Science and Technology, vol. 10, no. 9, pp. 81–91, 2024.

[6]. A. A. Mahindre and S. A. Kondekar, "Proactive health monitoring: Predictive analytics for early detection of diabetes risk," EasyChair Preprint, no. 15225, Oct. 2024.

[7]. A. Ahmed, J. Khan, M. Arsalan, K. Ahmed, A. A. Shahat, A. Alhalmi, and S. Naaz, "Machine learning algorithm-based prediction of diabetes among female population using PIMA dataset," Healthcare, vol. 13, no. 1, p. 37, 2025.

[8]. H. M. Deberneh and I. Kim, "Prediction of type 2 diabetes based on machine learning algorithm," International Journal of Environmental Research and Public Health, vol. 18, no. 6, p. 3317, 2021.

[9]. S. R. Mishra and S. Dash, "Predictive analysis on diabetes detection using Pima Indian diabetes dataset," International Journal of Research and Analytical Reviews, vol. 11, no. 2, 2024.

[10]. G. S and V. S. Reddy, "Type 2 diabetes mellitus: Early detection using machine learning classification," International Journal of Advanced Computer Science and Applications, vol. 14, no. 6, 2023.

[11]. S. Pandya, "Predicting diabetes mellitus in healthcare: A comparative analysis of machine learning algorithms," International Journal of Current Engineering and Technology, vol. 13, no. 6, pp. 545–546, Dec. 2023.

[12]. M. Al-Tawil, B. A. Mahafzah, A. Al Tawil, and I. Aljarah, "Bio-inspired machine learning approach to type 2 diabetes detection," Symmetry, vol. 15, no. 3, p. 764, Mar. 2023.

[13]. H. V. T. Huynh, L. Hui, N. H. Nguyen, and R. Qiao, "Performance analysis of diabetes detection using machine learning classifiers," International Journal of Management and Data Analytics, vol. 4, no. 1, pp. 43–54, Oct. 2024.

[14]. S. Mahajan, M. Rohra, P. K. Sarangi, and A. K. Sahoo, "Diabetes mellitus prediction using supervised machine learning techniques," in Proc. International Conference on Advancement in Computation & Computer Technologies (InCACCT), 2023.

[15]. B. Badeji-Ajisafe et al., "Early detection of diabetes using supervised learning approach," in IEEE Conference, 2023.

[16]. G. Tripathi and R. Kumar, "Early prediction of diabetes mellitus using machine learning," in IEEE Conference, 2023.

[17]. S. Mishra, V. A, and K. S, "Machine learning approaches for type-2 diabetes software predictor," in IEEE Conference, 2023.

[18]. H. O. Menge and P. Kuppuraj, "Machine learning-based early type 2 diabetes prediction," in Proc. International Conference on Emerging Research in Computational Science (ICERCS), 2024.

[19]. B. F. Wee, S. Sivakumar, K. H. Lim, W. K. Wong, and F. H. Juwono, "Diabetes detection based on machine learning and deep learning approaches," Multimedia

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr2270

Tools and Applications, vol. 83, pp. 24153–24185, 2024.
[20]. L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, "Early detection of type 2 diabetes mellitus using machine learning-based prediction models," Scientific Reports, vol. 10, p. 11981, 2020.
[21]. M. Matboli et al., "Machine learning-based stratification of prediabetes and type 2 diabetes progression," Diabetology & Metabolic Syndrome, vol. 17, p. 227, 2025.
[22]. C. H. Paparao et al., "Diabetes detection using machine learning," International Journal of Creative Research Thoughts, vol. 12, no. 5, May 2024.
[23]. P. Chowdhury, P. Barua, and M. N. Uddin, "Diabetes prediction using machine learning and hybrid deep learning ensemble technique," in Proc. IEEE Int. Conf. on Computing, Applications and Systems (COMPAS), 2024.
[24]. J. D. Akinyemi et al., "Machine learning-based diabetes risk prediction using associated behavioral features," Computational Open Journal, 2024.
[25]. K. C. Howlader et al., "Diabetes prediction using machine learning," Journal of Electrical Systems, vol. 20, no. 7, pp. 2244–2257, 2024.
[26]. G. Dharmarathne, "A novel machine learning approach for diagnosing diabetes using explainable AI," Healthcare Analytics, 2024.
[27]. B. Nguyen and Y. Zhang, "A comparative study of diabetes prediction based on lifestyle factors using machine learning," arXiv preprint, 2025.
[28]. M. Hasan and F. Yasmin, "Predicting diabetes using machine learning: A comparative study of classifiers," arXiv preprint, 2025.
[29]. P. B. Khokhar, C. Gravino, and F. Palomba, "Advances in artificial intelligence for diabetes prediction: Insights from a systematic literature review," arXiv preprint, 2024.
[30]. A. Hennebelle, H. Materwala, and L. Ismail, "HealthEdge: A machine learning-based smart healthcare framework for prediction of type 2 diabetes in IoT-edge-cloud systems," arXiv preprint, 2023