**SHAPLEY ADDITIVE EXTREME GRADIENT BOOSTING (SA-XGBOOST) BASED ON DIABETES PREDICTION**

**Abstract**

To improve patient survival and prevent major health problems, diabetes must be detected early. This research proposes a complete Machine Learning-based system for diabetes prediction using regular and locally acquired datasets. The system incorporates health-related characteristics including age, BMI, glucose level, blood pressure, and family history. To resolve this problem, the proposed a SHapley Additive eXtreme Gradient Boosting (SA-XGBoost) was constructed for the prediction of diabetes or non-diabetes. Furthermore, Z-Score Normalization Scaling Data (ZSNSD) collected data undergoes data preprocessing, including handling missing values, label encoding, and normalization. Then, key features are selected using Mutual Information and Recursive Feature Elimination (MIRFE) technique to enhance model focus and performance. Finally, the proposed SA-XGBoost model, the users to input their health parameters (such as AGE, BMI, BLOOD PRESSURE) and receive diabetes predictions instantly diabetes or not. Results prove that the system proposed performance metrics like accuracy, precision, recall, F1-score, this work gives reliable predictions with high accuracy 98% and strong potential for practical healthcare applications.

**Keywords:** Diabetes Prediction, Machine Learning, Feature Selection, XGBoost, SHAP, Health Monitoring.

**1. Introduction:**

Diabetes mellitus is an emerging international public health issue that impacts millions of people and poses heavy burdens to healthcare systems across the globe [1]. Early detection and prompt intervention are essential in halting the disease's progression and minimizing the risk of complications including cardiovascular disease neuropathy. Existing diagnostics depend on repeated medical checkups and laboratory diagnostics, which will not always be readily available and cost-effective in underserved communities [2]. Data-powered medicine and ML-based technologies that are now rising into prominence as options hold more potential by offering the possibility for prediction using inexpensive but quantifiable indicators of disease states [3]. Using traditional statistical techniques, ML models can perform pattern discovery on patient data. Using datasets like the PIMA Indian Diabetes Dataset, combined with locally retrieved health records, ML models can be trained to make reliable predictions of diabetes based on important features like glucose level, BMI, age, and family history [4]. Nevertheless, the accuracy of these models may be affected by problems like noisy data, redundant features, class imbalance, and lack of interpretability—factors that should be appropriately handled during model building [5].

This study proposes a robust ML-based diabetes prediction system. The system incorporates advanced preprocessing techniques for cleaning and normalizing the data, followed by feature selection using Mutual Information and RFE to focus on the most relevant predictors. To guarantee forecast transparency, SHAP and LIME are used to explain the classification model. XGBoost is one of several methods of classification that are assessed using cross-validation and common metrics.

**2. Literature Survey**

Over the years, numerous ML approaches have been proposed to improve the precision and effectiveness of diabetes prediction systems. Conventional diagnostic procedures rely on manual interpretation of patient data, which can lead to delays and inaccuracies, particularly in settings with limited resources [6]. ML presents an opportunity for automating this process and identifying intricate patterns in medical databases that conventional methods might not reveal. Several studies have used the PIMA dataset as a standard for testing and building prediction models because it has standardized features and is easily accessible [7].

The paper offers a nice comparison of different ML classifiers for predicting diabetes based on both accuracy and computational load. There are, however, some issues regarding model accuracy, possible overfitting, and missing information about the dataset, which can be addressed further in future research [8].

The article suggests a unique blend of ML methods for diabetes classification and prediction with encouraging outcomes (86.08% for MLP and 87.26% for LSTM). However, there are some issues regarding the accuracy, limitation of the dataset used, and hypothetical status of the IoT-based system. More experimentation, particularly on a diversified dataset and thorough real-world IoT integration, would be required to test the system's applicability in real healthcare environments [9].

This study provides a unique and highly structured method of postprandial blood glucose prediction through a gradient-boosting algorithm. It emphasizes data preprocessing, feature engineering, and model hyperparameter tuning. However, it requires further sharpening and proof, particularly under real-world scenarios, to conquer issues such as data accuracy, scalability, and accessibility [10].

This paper presents a realistic and practical perspective: In predicting undiagnosed T2DM, traditional regression models still hold their ground, especially in the early stages when data is limited. ML models like RF, LightGBM, and XGBoost did not show clinically significant improvement in accuracy. However, some (like LightGBM) were better regarding variable stability—a valuable property in long-term clinical use [11].

This research illustrates the application of comparing various ML method for predicting diabetes, with Random Forest being the most accurate one. It affirms that ensemble learning methods are stronger and yield better results than single models. This confines the power of interpretation of the accuracy improvement, though [12].

This research illustrates how Logistic Regression (LR) is used to identify the risk factors for diabetes disease based on p values and Odds Ratios (OR). While Random Forest performs well, it is still a black-box model. Clinical users prefer interpretable systems unless explainability (e.g., via SHAP values) is explicitly incorporated [13].

In order to predict diabetes based on a number of health characteristics in the dataset using supervised machine learning, this study compared two KNN algorithms. However, While the comparison is valid, more sophisticated algorithms (e.g., Random Forest, SVM, or Gradient Boosting) could provide better performance [14].

A comparative study between KNN and Naïve Bayes found NB to be more effective for diabetes prediction based on confusion matrix results. However, it lacked detailed accuracy metrics and dataset specifics. Finally, one notable research project used longitudinal Electronic Health Record (EHR) data to predict diabetes in the following year. After feature selection using chi-squared tests, and recursive elimination, they trained LR, XGBoost, and ensemble models [15].

This project tested 15 classifiers, and five of them were the main models: KNN, NB, and an ensemble approach. Nevertheless, this model is strong in terms of a blend of techniques and real-world application through a graphical interface and hence a deployable and reliable diagnostic tool [16].

The model integrates two powerful classifiers—SVM and ANN—to form a composite decision-making system that aims to improve accuracy. However, with the increasing global healthcare burden of diabetes, such data-driven tools could significantly impact patient survival rates and resource management [17].

In an experiment based on outpatient records at a Taipei Municipal medical center, scientists tackled Taiwan's growing burden of diabetes, which now besets more than 2.18 million Taiwanese. the fact underscores the robustness of ensemble methods to medical diagnostics, especially with good-quality clinical data [18]. The study identified RF and K-NN as the models, demonstrating superior predictive capability compared to other algorithms evaluated [19].

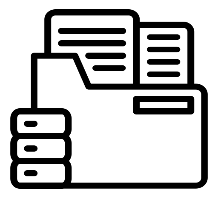
The common issue of class imbalance in medical datasets, the authors used SMOTE (Synthetic Minority Over-sampling Technique), which helped balance the data and improve model reliability. However, the study has some limitations. The size of the private dataset (only 203 individuals) and its specific demographic focus (female workers from a single region) restrict the generalizability of the model [20].

**3. Proposed Method**

This section proposed a SA-XGBoost model is interpreted for transparency using available interpretation packages such as SHAP. The methodology involves thorough data preprocessing and feature selection using Mutual Information and RFE to identify the most influential factors. Several classification algorithms are employed for model training, including Logistic Regression, KNN, SVM, and XGBoost. The models are evaluated using standard metrics with performance enhanced through hyperparameter tuning, cross-validation, and class balancing techniques like SMOTE and ADASYN.

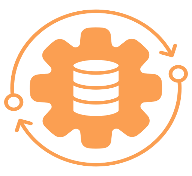
Data Collection

(Indian Diabetes Dataset)



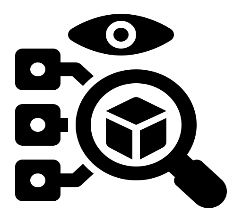
Data Preprocessing

Using ZSNSD



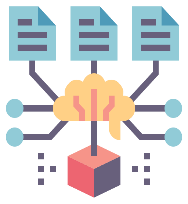
Feature Selection

(Mutual Information, RFE)

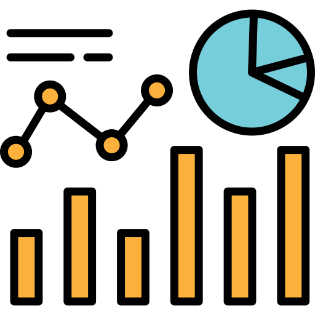


Model Building

SA-XGBoost

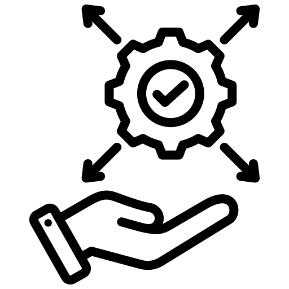


Model Evaluation



Handling Class Imbalance

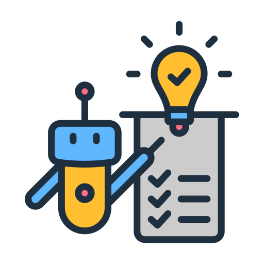
(SMOTE, ADASYN)



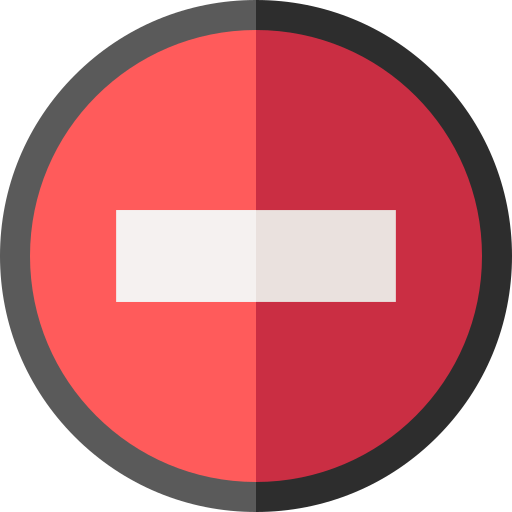
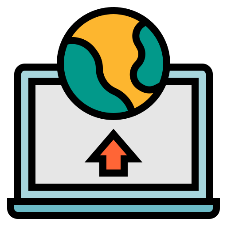
Model Explainability

(LIME, SHAP) Model Explainability

(LIME, SHAP)



Deployment



Diabetes

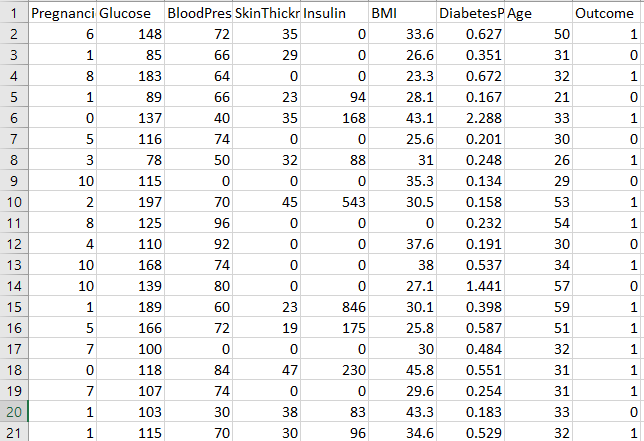
No Diabetes

**Figure 1: Architecture Diagram for Proposed Method**

The proposed architecture for the ML-based diabetes prediction system follows a streamlined flow from data acquisition to real-time prediction and feedback, as shown in Figure 1. Initially, health data is collected from the PIMA ID dataset and local sources, containing key features such as glucose level, BMI, blood pressure, and age. Always observation about the original data must be cleaned, with missing values addressed, and with normalization to be brought into an equal state. The subsequent step will comprise selecting important features using Mutual Information and Recursive Feature Elimination, which will help tighten the focus and keep the model's dimensionality low. The refined dataset is then split into training and testing sets, and several classification algorithms—SA-XGBoost —are trained. Each model is evaluated using accuracy, precision, recall, metrics. Class imbalance is addressed using SMOTE and ADASYN, and hyperparameters are optimized through Grid or Randomized Search. A feedback loop ensures continuous improvement, while future versions aim to integrate wearable sensor data and deep learning for broader applicability.

**3.1 Dataset Collection**

The first stage of this system entails collecting data from the PIMA Dataset and augmenting it with local patient data. The datasets comprise records with attributes such as glucose level, blood pressure, BMI, and a family history of diabetes. After the collection, data is cleaned by addressing missing values, outliers, and duplicates.



**Figure 2: The PIMA ID Dataset**

Figure 2 showed the PIMA ID Dataset; it contains 1000 records with 8 medical predictor variables and 1 target variable. The target variable, labelled as "Outcome," indicates whether a patient has diabetes (1) or not (0).

**3.2 Z-Score Normalization Scaling Data (ZSNSD)**

The proposedZSNSD system's core workflow begins with data collection, where diabetes-related medical records are gathered from sources like the PIMA Dataset and monitoring devices. The collected data undergoes data preprocessing, including handling missing values, label encoding, and normalization using Z-score scaling.

(1)

This equation 1 transformation converts the original feature values into a distribution with a mean of 0 and a standard 1.

To bring all features to a uniform scale and avoid bias during training, min-max normalization is applied to each feature. This scales input values to a range between 0 and 1. The formula is given equation 2:

(2)

**3.3 Mutual Information- Recursive Feature Elimination (MI-RFE)**

The robust method for selecting the most relevant features before training a ML model, a filter method for assessing how much a feature contributes to the output label, and a wrapper method that eliminates less important features recursively using a trained model.

Following equation 3 normalization, feature selection is conducted using two primary techniques. The first is Mutual Information (MI), which quantifies the dependency between each feature J and the target class K. A high MI score indicates that the feature contributes significantly to diabetes prediction:

(3)

The equation 4 method, RFE ranks and removes the least important features based on model performance iteratively until the optimal feature set is achieved.

(4)

Optimizes its predictions using the following regularized loss equation 5: Here, is the loss method, ​ is the number of leaves, and are regularization terms to prevent overfitting.

(5)

Once feature selection is complete, the dataset is split into training and testing sets, often in an 80:20 ratio. One of the baseline models is MI-RFE, which estimates the probability of diabetes through a sigmoid function. This function converts inputs into a number between 0 and 1, representing the likelihood of having diabetes.

**3.4 SHapley Additive eXtreme Gradient Boosting (SA-XGBoost)**

Once the best-performing model is selected (typically XGBoost), it's essential to ensure that predictions are interpretable, especially for medical use. Two explainability tools are integrated: SHAP (SHapley Additive exPlanations), which explains each prediction by computing the contribution of each feature.

XGBoost creates a group of option trees, where the prediction ​ for an input instance is given by the sum of the predictions from all trees in the equation 6:

(6)

Each tree makes a prediction based on a set of features, and these trees are combined to produce the final output.

Shapley values provide a fair allocation of the "credit" (or prediction) to each feature based on its contribution. For a given model, the Shapley value for a feature is computed by considering all possible permutations of the features and averaging the feature j's marginal contributions, The Shapley value for new j in a model’s prediction y​ can be equation 7:

(7)

This formula determines feature j's average contribution over all potential feature subsets, considering the difference in model predictions when feature is added to each subset.

The equation 8 shows that the final value is a combination of the baseline find and the contributions from each feature, which are computed using the Shapley values.

(8)

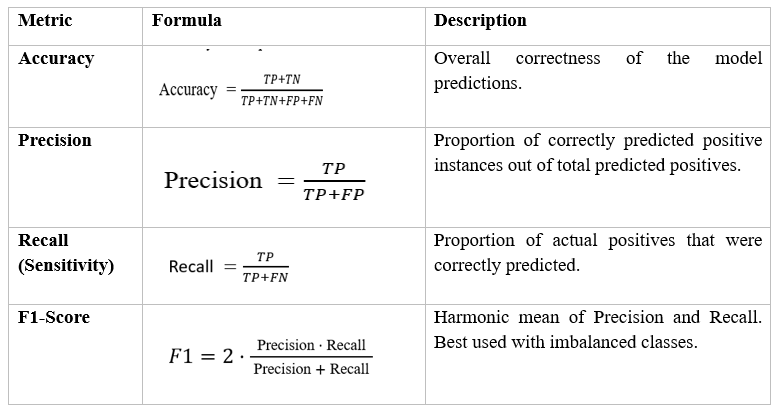
Finally, the trained model is deployed through either a web interface or Android app. Users provide personal health inputs (like age, BMI, and glucose), upon which they receive real-time predictions for utilizing the model. Thus, the model's real-world preventive healthcare applicability is also realized.

**4. Result and Discussion**

Based on performance criteria such as accuracy, precision, recall, and F1-score, this section presents the findings and evaluation of the proposed machine learning (ML)-based diabetes prediction model. The experiments were performed on the PIMA dataset and local health data from a textile factory in Bangladesh. The system's performance was tested on various ML models, with particular emphasis given to XGBoost, which was found to be the most accurate and consistent model. Class balancing (SMOTE and ADASYN), hyperparameter tuning (Grid Search and Randomized Search), and model interpretability using LIME and SHAP were also used for evaluation.

**4.1 Model Evaluation and Optimization**

After modelling phase, models are evaluated using confusion matrix in which number of TP, TN, FP and FN can be found. Several performance metrics can be derived from this matrix like Accuracy which indicates ratio of correct predictions:

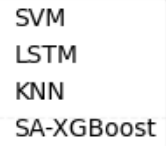
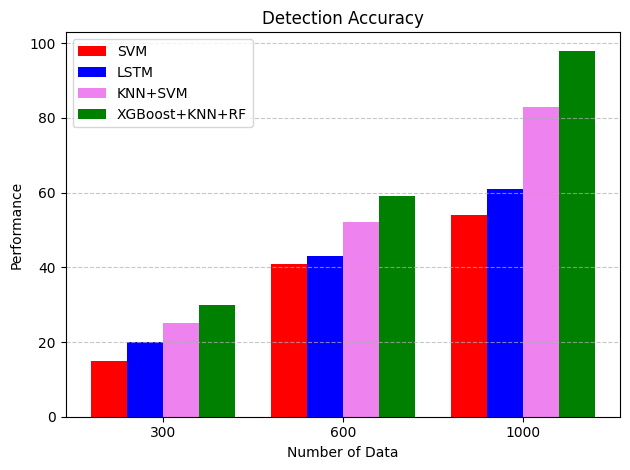


The Table 1 Hyperparameter tuning is performed using Grid Search or Randomized Search techniques to further improve model performance. Class imbalance—where diabetic patients are fewer in number—is addressed using SMOTE. The balanced dataset is then used to retrain the classifiers for improved recall and generalization.

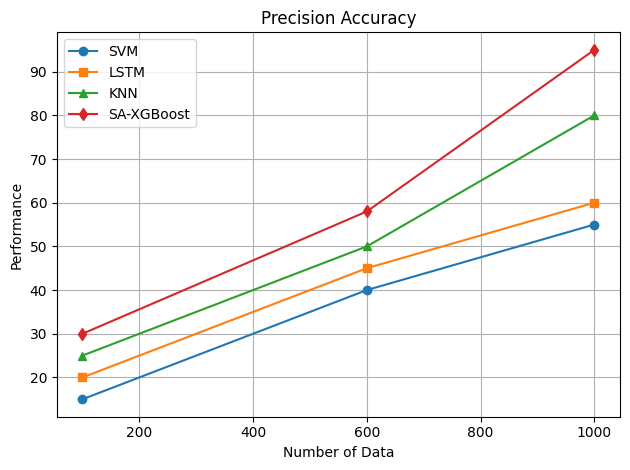
**Table 2. Experiment Setup**

|  |  |
| --- | --- |
| Parameter | Value/Description |
| Dataset | PIMA Indian Diabetes + Local Data |
| Number of Samples | 1000 |
| Features Used | Age, BMI, Glucose, Insulin, etc. |
| Feature Selection | Mutual Information, RFE |
| Data Balancing | SMOTE, ADASYN |
| Hyperparameter Tuning | Grid Search, Randomized Search |
| Explainability Tools | Jupyter |

Table 2 shows the experimental configuration used for testing the diabetes prediction model. Data preprocessing steps included removing missing values ​​and cleaning the outlier. After feature selection through Mutual Information and RFE, the dataset was divided into training and test sets in an 80:20 ratio and tested using cross-validation to guarantee robustness.

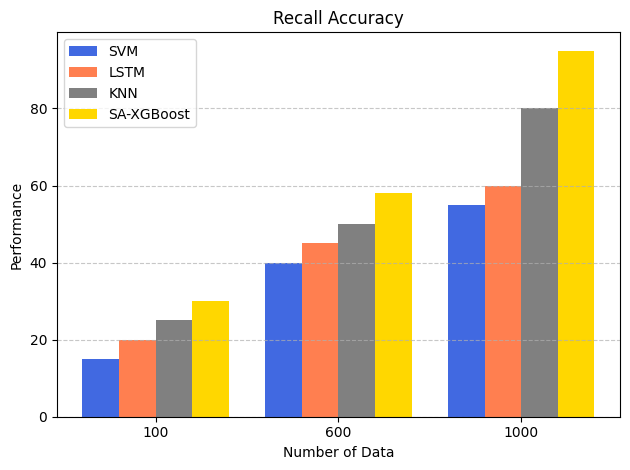


**Figure 3: Detection Accuracy**

Figure 3 illustrates the comparative detection accuracy of ML models—LSTM, SVM, and the proposed hybrid model combining KNN, and Random Forest—on varying data sizes (300, 600, and 1000 records). As the dataset size increases, a clear upward trend in prediction performance is observed across all models. However, the proposed hybrid ensemble approach significantly outperforms the others at every stage. At 300 data samples, SA-XGBoost achieves approximately 30% accuracy, while the SVM, LSTM, and KNN models lag at 15%, 20%, and 25%, respectively. This performance gap widens as the data volume increases: at 1000 samples, the proposed method achieves nearly 98% accuracy, whereas KNN, LSTM, and SVM reach only about 84%, 62%, and 55%, respectively. These results of the proposed ensemble method in learning from complex patterns and feature interactions present in medical datasets like those used for diabetes prediction.

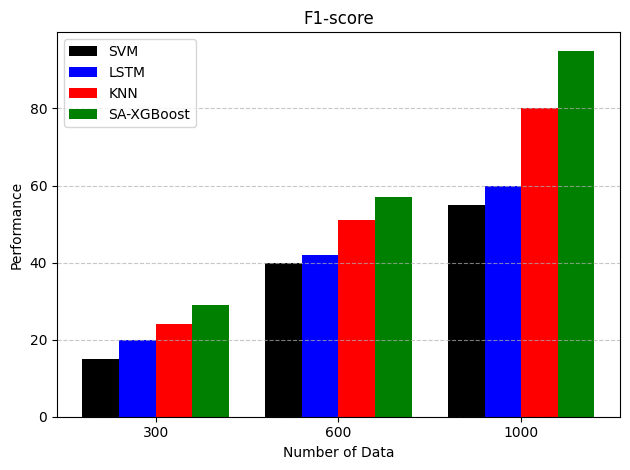
**Figure 4:** **Precision Accuracy**

Figure 4 presents a comparative analysis of precision accuracy among four models—SVM, LSTM, KNN, and the proposed hybrid model SA-XGBoost —across different dataset sizes (150, 600, and 1000 data points). At 150 data samples, the proposed hybrid model begins with a precision accuracy of about 30, compared to SVM (15), LSTM (20), and KNN (25). At 1000 samples, SA-XGBoost achieves the highest precision accuracy, nearing 95, while KNN, LSTM, and SVM reach approximately 80, 60, and 55, respectively. This upward trend confirms that the hybrid model scales effectively with increasing data size and maintains superior performance in predicting diabetic outcomes with high precision.



**Figure 5: Recall Accuracy**

Figure 5 illustrates a comparative evaluation of recall accuracy across four models: SVM, LSTM, KNN, and the proposed hybrid SA-XGBoost, as the dataset size increases from 100 to 1000 samples. Recall accuracy, which measures the proportion of actual positive cases (i.e., diabetic cases) correctly identified, is critical in healthcare diagnostics to ensure high sensitivity and reduce missed diagnoses. At 100 data samples, SA-XGBoost starts with the highest recall (around 30%), compared to SVM (15%), LSTM (20%), and KNN (25%). As the dataset grows to 600 samples, the hybrid model continues to outperform others, reaching approximately 58% recall, while SA-XGBoost, and SVM follow behind with 50%, 45%, and 40%, respectively. At the 1000-sample mark, SA-XGBoost achieves a peak recall of 90%, while KNN, LSTM, and SVM attain 80%, 60%, and 55%, respectively. These results affirm the proposed hybrid model's superior sensitivity and ability to detect diabetic cases with high reliability correctly.



**Figure 6: F1- Score**

Figure 6 demonstrates a comparative analysis of the F1 Score—an important metric that balances precision and recall—across four models: SVM, LSTM, KNN, and the proposed hybrid SA-XGBoost, evaluated at increasing dataset sizes (300, 600, and 1000 samples). The F1 Score is particularly vital in medical diagnostics, such as diabetes prediction, where both FP and FN can have serious consequences. At 300 data samples, the F1-scores for SVM, LSTM, KNN, and SA-XGBoost were approximately 15%, 20%, 25%, and 30% respectively. As the dataset grew to 600 samples, the hybrid model's F1-score reached nearly 57%, while other models followed behind—KNN at 51%, LSTM at 43%, and SVM at 40%. At 1000 samples, SA-XGBoost achieved a 92%, compared to KNN (80%), LSTM (60%), and SVM (55%). The superior F1-score of the proposed hybrid model validates its overall robustness and effectiveness in diabetes prediction, especially in scenarios where sensitivity and specificity are crucial.

**5. Conclusion**

This work introduces a ML-powered structure to predict early and accurate diabetes, applies a multi-model strategy to increase classification accuracy and reliability. The proposed system employs the most appropriate health indicators and employs advanced feature selection methods like RFE. Several classifiers, XGBoost, were tested, and the best performance was obtained using the hybrid ensemble of SA-XGBoost on all metrics of evaluation. Simultaneously, model interpretability was attained by applying SA-  
XGBoost method facilitating interpretable interpretation of model selections. This helps clinicians and medical practitioners to rely and comprehend the outputs of the model. The results of the experiment indicate that the ensemble model correctly identifies diabetes with high accuracy and recall and has an F1-score over 90% when the size of the dataset increases. The method reduces misclassifications to a minimum and provides balanced sensitivity and specificity, which are critical in medical diagnostics. Future plans involve integrating wearable sensor data as well as deep learning models in order to enhance real-time prediction and monitoring potential even further. The framework developed provides a scalable, interpretable, and solid solution for diabetes detection at the early stage, allowing proactive health care and enhanced patient outcomes.

**Reference:**

1. Hasan, Md Kamrul, et al. "Diabetes prediction using ensembling of different machine learning classifiers." IEEE Access 8 (2020): 76516-76531.
2. Ahmed, Usama, et al. "Prediction of diabetes empowered with fused machine learning." IEEE Access 10 (2022): 8529-8538.
3. Xie, Jinyu, and Qian Wang. "Benchmarking machine learning algorithms on blood glucose prediction for type I diabetes in comparison with classical time-series models." IEEE Transactions on Biomedical Engineering 67.11 (2020): 3101-3124.
4. Chauhan, Tannu, et al. "Supervised and unsupervised machine learning based review on diabetes care." 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS). Vol. 1. IEEE, 2021.
5. Fazakis, Nikos, et al. "Machine learning tools for long-term type 2 diabetes risk prediction." ieee Access 9 (2021): 103737-103757.
6. Sakib, Shadman, et al. "Performance analysis of machine learning approaches in diabetes prediction." 2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC). IEEE, 2021.
7. Ahmed, Nazin, et al. "Machine learning based diabetes prediction and development of smart web application." International Journal of Cognitive Computing in Engineering 2 (2021): 229-241.
8. Lyngdoh, Arwatki Chen, Nurul Amin Choudhury, and Soumen Moulik. "Diabetes disease prediction using machine learning algorithms." 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES). IEEE, 2021.
9. Butt, Umair Muneer, et al. "Machine learning based diabetes classification and prediction for healthcare applications." Journal of healthcare engineering 2021.1 (2021): 9930985.
10. Pustozerov, Evgenii A., et al. "Machine learning approach for postprandial blood glucose prediction in gestational diabetes mellitus." Ieee Access 8 (2020): 219308-219321.
11. Kopitar, Leon, et al. "Early detection of type 2 diabetes mellitus using machine learning-based prediction models." Scientific reports 10.1 (2020): 11981.
12. Soni, Mitushi, and Sunita Varma. "Diabetes prediction using machine learning techniques." International Journal of Engineering Research & Technology (IJERT) 9.09 (2020): 2278-0181.
13. Maniruzzaman, Md, et al. "Classification and prediction of diabetes disease using machine learning paradigm." Health information science and systems 8 (2020): 1-14.
14. Febrian, Muhammad Exell, et al. "Diabetes prediction using supervised machine learning." Procedia Computer Science 216 (2023): 21-30.
15. Deberneh, Henock M., and Intaek Kim. "Prediction of type 2 diabetes based on machine learning algorithm." International journal of environmental research and public health 18.6 (2021): 3317.
16. Sarwar, Abid, et al. "Diagnosis of diabetes type-II using hybrid machine learning based ensemble model." International Journal of Information Technology 12 (2020): 419-428.
17. Nadeem, Muhammad Waqas, et al. "A fusion-based machine learning approach for the prediction of the onset of diabetes." Healthcare. Vol. 9. No. 10. MDPI, 2021.
18. Chou, Chun-Yang, Ding-Yang Hsu, and Chun-Hung Chou. "Predicting the onset of diabetes with machine learning methods." Journal of Personalized Medicine 13.3 (2023): 406.
19. Dritsas, Elias, and Maria Trigka. "Data-driven machine-learning methods for diabetes risk prediction." Sensors 22.14 (2022): 5304.
20. Tasin, Isfafuzzaman, et al. "Diabetes prediction using machine learning and explainable AI techniques." Healthcare technology letters 10.1-2 (2023): 1-10.