ETL-POXGB: A NOVEL CLASSIFICATION FRAMEWORK COMBINING ETL DATA INTEGRATION, ENSEMBLE FEATURE SELECTION, AND PSO OPTIMIZATION

V. USHA[1], DR. N. R RAJALAKSHMI [2]

[1]Research Scholar, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Computer Science and Engineering, Tamil Nadu, India
[2] Professor, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Computer Science and Engineering, Tamil Nadu, India
E-mail: [1]v.ushavelusamy@gmail.com, [2]drnrrajalakshmi@veltech.edu.in

ABSTRACT

The complicated metabolic illness known as diabetes has multiple causes, including genetics, the environment, and lifestyle choices. It is characterized by persistently elevated blood sugar levels. Therefore, to reduce its harmful repercussions, early detection of diabetes is crucial. The growing integration of Information Technology (IT) in predictive healthcare analytics assists in developing more accurate, scalable disease prediction models. IT enhances the current research by employing optimization methodologies, data preparation methods, and machine learning algorithms to increase the accuracy of diabetes predictions. The current research presents an optimization strategy for improving diabetes prediction by combining multiple feature selection models. This work introduces a novel Ensemble Fisher score Kolmogorov-Smirnov score and Chi-Square (FKCS) model that effectively improves forecast accuracy and efficiency. To test machine learning algorithms for predicting diabetes, diabetic datasets like Pima, Iraq, and Frankfurt were used. These datasets came from different sources and had important clinical characteristics. The findings were analysed using multiple statistical machine-learning measures and a stratified cross-validation approach. Among all classifiers, the highest level of accomplishment was demonstrated by the Extract Transform Load: Particle Swarm Optimization XGBoost (ETL-POXGB), achieving an impressive accuracy percentage of 97.16%. The model was validated using Precision, Recall, F1 Score, ROC AUC, CK, and MCCoeff on the merged dataset. In all aspects of evaluation, superior performance was displayed by our proposed model.

Keywords: *Multimodal Medical Data, Optimization, Integrated Diabetic Datasets, Machine Learning, Cross-Validation, ETL-POXGB.*

1. INTRODUCTION

A metabolic illness defined by persistently high blood sugar levels, diabetes mellitus is more popularly known as diabetes. In medical parlance, diabetes mellitus is known simply as "diabetes." This disorder develops when the body stops making or properly using insulin, a hormone that controls blood sugar levels. Different factors contribute to the development of the three primary forms of diabetes. When a person develops type 1 diabetes, their immune system targets and destroys insulin-producing beta cells in the pancreas. Although researchers have not pinpointed a single cause for this autoimmune response, they do believe that both genetic and environmental factors contribute. For the rest of their lives, people with type 1 diabetes must take insulin to keep their blood sugar levels under control.

The most prevalent kind of diabetes, type 2, occurs when the body stops responding normally to insulin or when insulin production drops below normal levels. Type 2 diabetes can run in families and is exacerbated by factors such as obesity, inactivity, poor nutrition, and heredity. A balanced diet, frequent exercise, and, in rare instances, medication can help keep this form of diabetes under control. The failure of the body to fabricate adequate insulin to fulfil the increased demands during pregnancy causes gestational diabetes. Both hereditary and hormonal factors can influence gestational diabetes. Women who have diabetes while pregnant are more likely to develop type 2 diabetes later in life, although it usually goes away after birth.

Hyperglycemia, or diabetes, is a medical term for a collection of metabolic disorders. One of the top killers globally in the last decade or so is diabetes. In twenty nineteen, diabetes was responsible for the deaths of approximately one and a half million individuals globally, as reported by the World Health Organization. Worldwide, around 537 million persons (ranging in age from 20 to 79) were dealing with diabetes in twenty twenty-three, as reported by the International Diabetes Federation (IDF). Figure 1 illustrates type 2 diabetes and its potential serious implications.

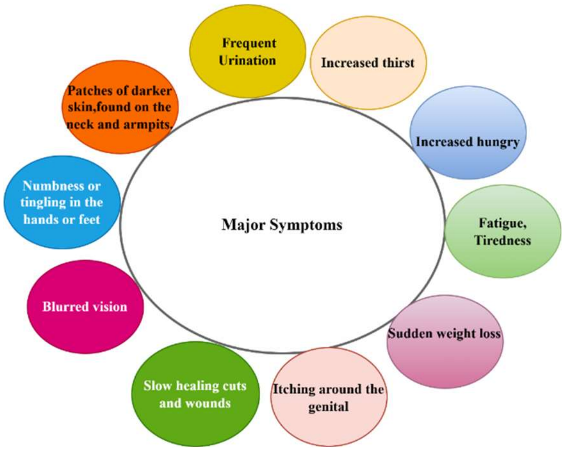

Frequent Urination
Increased thirst
Patches of darker skin,found on the neck and armpits
Increased hungry
Numbness or tingling in the hands or feet
Major Symptoms
Fatigue, Tiredness
Blurred vision
Sudden weight loss
Slow healing cuts and wounds
Itching around the genital

Figure 1: Diabetes' Symptoms

As an Artificial Intelligence (AI)-based medical application, Machine Learning (ML) aims to improve Computer-Aided Diagnosis (CAD) systems' ability to anticipate diabetes-related parameters. ML makes use of analytical models that can learn on their own from data, finding patterns and making decisions quickly. Just like humans, ML can learn to overcome its shortcomings with experience, making it an invaluable tool for solving complicated problems.

The medical field considers ML-based algorithms crucial, especially for improving CAD schemes that predict diabetes-related variables and survival rates. Many areas of computer vision and medicine, including radiology, demonstrate the indispensable nature of ML and its adaptability. With this technology, may better diagnose deadly diseases, as well as streamline administrative operations in hospitals, better map and treat infectious diseases, and personalize medical therapies.

Getting useful information from raw data and turning it into features is the goal of quality business in diabetes prediction. This will help the model find risk-related patterns and correlations better. To guarantee that the designed characteristics are in line with clinical findings, this approach necessitates a thorough grasp of the subject and close cooperation with healthcare professionals. Our objective is to equip the machine learning models with a thorough and accurate set of attributes that will enable them to generate precise predictions regarding diabetes.

These diverse investigations have greatly improved the accuracy of insulin forecasting. Nevertheless, these conclusions are derived from a solitary, limited dataset. Given the multiple diabetes statistics available and the insufficient data included in one set of data, it is recommended to combine various diverse information to improve the accuracy of diabetic disease predictions. It is recommended to combine various diverse information to progress the exactness of diabetic disease predictions. The merging of datasets is itself a feature fusion process, so keep this in mind throughout. When you mix datasets from diverse sources, the features from those sources are going to merge, even though they represent distinct kinds of material. A lot of databases are multi-faceted when it comes to feature fusing.

Disease prediction has been transformed by the incorporation of information technology (IT) into healthcare analytics, which has made it possible for intelligent decision-making systems. The accuracy and practicality of traditional diabetes prediction models are generally limited by inadequate feature selection, ineffective classification algorithms, and poor data preprocessing. This research addresses these challenges by introducing a machine-learning framework that enhances data quality, feature selection, and classification performance.

Following is a short overview of the paper's key findings:

1. Imputed the missing values in Three different datasets using K-Nearest Neighbors imputation to reduce the bias and increase the data quality.

2. Investigated and enhanced diabetes prediction by merging three diverse datasets from various sources and converting them into a single data set using the ETL (Extract, Transform, and Load) data combination Technique.

3. After the ETL process, hot deck imputation was performed on the combined dataset to improve model performance.

4. Utilized ensemble methods (Fisher score, Kolmogorov-Smirnov score, and Chi-Square) Ensemble FKCS to select important features for diabetic classification.

5. Implemented PSOptimization to optimally select the best classifier for prediction

Here's the outline of the paper workflow. In Section 2, the relevant word of diabetes prediction is delved into, while Section 3 offers a full overview of the system's design and an introduction to the activities of the individual modules. The dataset description, experimental settings, and more are provided. In Section 4, the experimental assessment metrics, experiment findings, and how to analyze the results are discussed. The paper concludes in Section 5.

2. LITERATURE REVIEW

To forecast the likelihood of hypoglycemia within a 24-hour timeframe, a novel method for making such predictions [1] integrated data fusion with classifier consensus. A total of 54 patients' diabetes-related records from the UC Irvine dataset are utilized by the approach. Despite limited data and the use of self-monitoring blood glucose, the results are encouraging. When compared to other studies in the same field, this method improves hypoglycemia prediction while producing fewer false alarms.

Type 1 and Type 2 diabetes, as well as gestational diabetes, are the focus of the danger forecast models and algorithms studied in [2]. The Pima Indian Diabetes Dataset was used to build Machine Learning models, which were classified using Gradient Boost, Decision Tree, and Logistic Regression techniques. Decision Tree ranked first in memory, accuracy, precision, and Fi score.

Classifying diabetes in Iraq using medical tests and bodily traits is the subject of the study, [3] which employs a Long-Short Term Memory (LSTM) neural network. Finding five of eleven indicators to be most important, the study was able to reduce the number of features employed and the expense of yearly check-ups while still offering a classification accuracy of up to 98% among people with diabetes, those without diabetes, and those at risk of developing the disease.

If identified early, diabetes, a chronic condition, can be effectively controlled. Automated systems for diabetes patient detection have been developed [4] using machine learning approaches, such as ontology-based machine learning. Support vector machines (SVMs), k-nearest neighbors (KNNs), artificial neural networks (ANNs), decision trees, logistic regression, naive bayes, and other common ML approaches are reviewed in this study. When comparing accuracy, ontology classifiers and support vector machines (SVM) come out on top. Numerous domains, including healthcare, have demonstrated the usefulness of the Semantic Web—which encompasses ontology. To improve prediction, recommendation, and decision-making, the project invites researchers to offer novel ideas.

The early detection of type 2 diabetes mellitus (T2DM) is crucial in the prevention and mitigation of complications associated with this chronic condition. The modified [5] mayfly-support vector machine was used to create a multi-class predictive model that makes use of machine learning algorithms. Maximum test accuracy of 94.5% was achieved when the model was tested using a benchmark PIMA dataset and local hospitals. Both theoretical analysis and practical trials were used to determine how well the model worked. Improved T2DM prediction performance using the modified Mayfly-SVM method paves the way for joint efforts between patients and doctors to reduce the risk of complications. Developing practical health recommendations for populations at high risk should guide future research.

A metabolic disorder, diabetes mellitus is defined by persistently elevated blood sugar levels. If left untreated, it might lead to consequences. The purpose of this work [6] is to build a radial basis function and Bayesian network-based mixed Ensemble Learning (EL) system for making predictions. With a precision of 97.11%, the EL method surpasses five different machine-learning approaches. Diabetes experts could be able to better diagnose patients and prescribe effective treatments with the use of this ensemble learning.

Worldwide, 425 million individuals are living with diabetes, a metabolic disorder that causes insulin resistance. Important tools for early detection include artificial intelligence and data mining techniques. To identify those who have diabetes, offer a random forest algorithm that has been fine-tuned using the best parameters (RFWBP) [7]. After receiving its training from more traditional means, the algorithm employs data processing techniques and features engineering. With and without fold-up cross-validation, the RFWBP obtains an accuracy of 90.68% and 95.83%, respectively. Compared to traditional machine learning methods, the RFWBP performs far better, according to the experimental data.

Disabilities and early deaths caused by diabetes are significant global health concerns. For diabetes prediction, this [8] research introduces a medical decision system that makes effective use of Deep Neural Networks (DNN). In healthcare, these sophisticated algorithms can improve transparency, adaptability, and decision-making efficacy. Reducing healthcare service costs and enhancing decision accuracy, the proposed method attained an accuracy of 99.75% and an F1-score of 99.66% when compared to machine learning techniques.

One weighted k-nearest neighbor classification approach is the standard deviation K-nearest neighbor (SDKNN), which measures the distance between the training and testing datasets by utilizing the standard deviations of characteristics. This approach [9] calculated distance differently depending on the standard deviation of characteristics, which increases classification accuracy. The method outperformed traditional methods with an average classification accuracy of 83.2% while applied toward the Pima Indian Diabetes Dataset (PIDD).

In this paper, [10] researchers use the Pima diabetes dataset and five different boosting algorithms to investigate latent machine knowledge in healthcare for diabetes prediction. Gradient boosting outperformed all other classifiers into provisions of exactness, reaching 92.85%. This approach applies to other diseases with comparable predicate indications because it outperformed recent research in prediction accuracy.

Individual of the foremost causes of diseases among non-communicable diseases, diabetes impacts 537 million people worldwide. Excessive weight, cholesterol, a genetic predisposition, a sedentary lifestyle, and unhealthy eating habits are the root causes. Using a proprietary dataset from Bangladesh and various machine learning algorithms, a system for autonomous diabetes prediction has been created. Several ensemble methods, a semi-supervised model [11] using extreme gradient boosting, the SMOTE (Synthetic_ Minority Over-sampling Technique) and ADASYN (Adaptive_Synthetic) methodologies, and mutual information feature selection are all part of the system. With an area under the curve (AUC) of 0.84, the system executed XGBoost with 81% accuracy. Additionally, the system incorporates an approach to explainable AI and a domain adaption mechanism.

Accurately predicting diabetes utilizing heterogeneous data sources is critical due to the increasing number of people with diabetes. To solve this problem, missing values in a fused dataset are filled in using a graph representation approach and repeated imputation. Both the training and prediction processes make use of the logistic regression model and stacking method. The fusion approach [12], when applied to two different datasets, improves the accuracy of diabetes predictions. The World Medical Health dataset is best served by a combination of the Genetic_Algorithm_Robust_Artificial_Neural_Net work_Ensemble (GRAPE) and stacking models, whereas the Pima dataset is best served by a combination of the MICE and stacking models. Any situation using the same label types and various feature attributes can be addressed using this approach.

The research [13] delves deeply into employing machine learning algorithms for diabetic prognosis, concentrating specifically on the multilayered Data Amalgamation and Fusion Enabled Diabetes Prediction using Machine Learning (DPEMDFML) prototype. This innovative model synergistically joins both Artificial Neural Networks alongside Support Vector Machines to excavate nuanced insights from voluminous datasets, achieving a stunningly accurate prediction rate exceeding 97%. Nonetheless, additional testing and evaluation of this model's exact predictive capabilities is still necessary to completely comprehend its full scope for revolutionizing individualized medical care.

This author [14] expounds on a novel hybrid optimization strategy for diabetes prediction leveraging machine learning techniques. Two datasets from the renowned Pima Indian diabetes database were enlisted alongside a feature selection process. The proposed technique, termed the Binary Grey Wolf-Crow Search Optimization, amalgamates the principles of Binary Grey Wolf Optimization and Crow Search Optimization. Simulation results demonstrated the method surpasses prevailing approaches, achieving an impressive accuracy of 96.62%. Furthermore, the approach optimizes the quantity of concealed neurons within Support Vector Machines by refining the number of hidden nodes. Looking ahead, this hybrid strategy shows promise for practical deployment in healthcare settings to enhance disease risk assessment and patient outcomes.

This paper proposed [15] developing a novel machine learning form for predicting with detecting microvascular complications among Type 2 diabetic patients. The suggested model utilizes an enhanced version of the coati algorithm for feature selection before comparing its diagnostic capabilities against more conventional classifiers such as extra gradient boosting, K-nearest neighbors, support vector machines, random forests, AdaBoost, decision trees, and artificial neural networks. This diagnostic tool classifies both diabetic retinopathy and diabetic neuropathy based on factors including age, sex, body mass index, blood pressure readings, foot pulse scans, family history of diabetes, and medication adherence patterns. Testing revealed AdaBoost to be the optimal estimator for identifying diabetic retinopathy and neuropathy, achieving remarkably high accuracy levels of 99.9% and 94.78%, respectively, according to our findings.

An investigation [16] provided there builds a machine learning framework for forecasting diabetes utilizing a combination of Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs). To decide whether the diagnosis is positive or negative, the model uses a 70:30 split of training and testing data. Using a patient's real-time medical information, the combined model outperforms prior methods in accurately projecting their diabetic state, with a prediction score of 94.87.

A stacking-based integrated Kernel-Enhanced Learning Machine (KELM) model to envisage the possibility of Type-II diabetes within five existence of evaluation is proposed in this work. This model [17] applies the Hybrid_Algorithm for Fuzzy_Particle_Swarm Optimization (HAFPSO) algorithm and Min-Max normalization to two datasets: One from the Diabetic Research Center and the other from the Pima Indian Diabetic Database. By integrating the predictions of twenty support learners, the sculpt is skilled via KELM as a meta-classifier. The technique has a very high accuracy rate of 98.5% when tested for precision, specificity, sensitivity, and mathematical correlation coefficient.

Common and incurable, diabetes is sometimes referred to as the "second killer" disease. This [18] research suggests a way to diagnose diabetes early on by combining ensemble learning with Boruta feature selection. For classification and unsupervised clustering, the model employs ensemble learning and the K-Means++ algorithm. When tested on the PIMA mellitus dataset, the model performed brilliantly, scoring 98% accuracy. This model outperforms the others when it comes to diabetes prediction and performance, according to the results.

The "second killer" disease, diabetes, is a chronic illness for which there is now no treatment. This research [19] suggests a way to diagnose diabetes early on by combining ensemble learning with Boruta feature selection. The model achieved a 98% accuracy rate when tested on the PIMA Indian diabetes dataset. The results show that this model outperforms other models when it comes to diabetes prediction, which emphasizes the need for early intervention and better health outcomes.

The investigation [20] uses the Iraqi Patient Dataset of Diabetes to propose a multi-classification architecture based on pipelines for diabetes prediction. To solve challenges like as imbalanced datasets, missing values, and a lack of labeled data, the system employs a variety of machine learning models and pre-processing methods. This model outperforms prior models in terms of accuracy, precision, recall, F1 score, and area under the curve (AUC). Despite the dataset's imbalance, the framework shows promise in accurately predicting diabetes in Iraqi diabetic patients, indicating the potential of machine learning techniques for diabetes management. Additional research is needed to improve the framework's applicability to additional datasets and demographics.

3. PROPOSED FRAMEWORK

Figure 2, which depicts the system architecture, illustrates the primary procedure to integrate data sources and forecast diabetes. Input of data, manipulation of data, and finally, prediction make up the three main phases of the system. The diabetes prediction result is generated by taking three different datasets as input, merging them into a single fused dataset, and then training and predicting from this combined dataset. Below, you will find a description of these three stages.

3.1 Dataset Insight

Three separate datasets were used in this investigation. PIMA, Iraqi, and Frankfurt are the three flavors. Also included in the PIMA dataset are nine columns, eight independent variables that are “Gender”, “Pregnant”, “Glucose”, “Diastolic_BP”, “Skin_Fold”, “Serum_Insulin”, “BMI”, “Diabetes_Pedigree”, “Age”, and one outcome parameter.

Most of the people who took part in the study were women. There is a record of 768 observations. Furthermore, 268 instances have been confirmed as positive and 500 cases as negative. Data was collected from the laboratories of Medical City Hospital and the Specialized Center for Endocrinology and Diabetes-Al-Kindy Teaching Hospital, information was also collected from Iraqi individuals. In other words, the data came from medical facilities.

Patient records were obtained, processed, and subsequently entered into a database to form the diabetes dataset. A total of 1000 patients were seen, comprising 436 females and 565 males. Total 14 attributes not including dependent variable. Frankfurt has 2000 individuals concerned. There was a total of eight independent variables one dependent variable there.

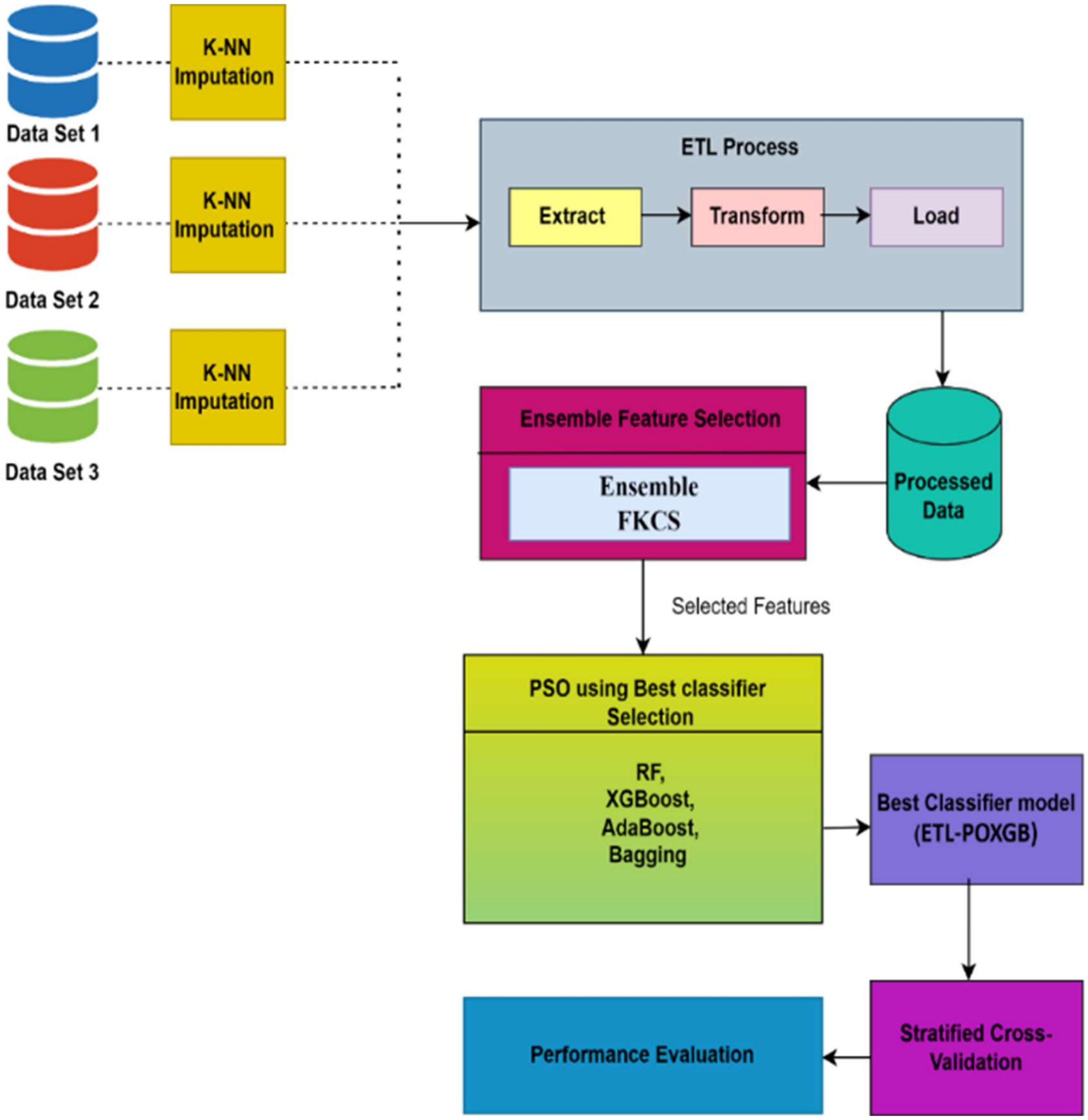


Figure 2: Proposed System Framework

Most of the people who took part in the study were women. There is a record of 768 observations. Furthermore, 268 instances have been confirmed as positive and 500 cases as negative. Data was collected from the laboratories of Medical City Hospital and the Specialized Center for Endocrinology and Diabetes-Al-Kindy Teaching Hospital, information was also collected from Iraqi individuals. In other words, the data came from medical facilities. Patient records were obtained, processed, and subsequently entered into a database to form the diabetes dataset. A total of 1000 patients were seen, comprising 436 females and 565 males. Total 14 attributes not including dependent variable. Frankfurt has 2000 individuals concerned. There was a total of eight independent variables one dependent variable there.

3.2 Data Preprocessing

It's taking raw data and turning it into something usable and useful. A quick rundown of the preprocessing steps in the suggested framework is as follows: eliminating duplicate samples, Feature Harmonization**,** Handling empty values, and trait selection from feature lists.

3.2.1 Feature harmonization

Feature harmonization intends to improve computational efficiency, reduce errors, and provide clarity and consistency across several datasets. Following this process simplifies data integration, improves model performance, and makes collaboration and maintenance much easier. Advantages include increased precision in data processing, less need for human involvement, and improved readability.

In this research work, the PIMA dataset is used to ensure that all subsequent datasets have matching feature names. To make things clearer and more consistent with industry norms, and also renamed a few attributes in the Frankfurt dataset. Notably, 'Pregnancies' has been renamed 'Pregnant', 'BloodPressure' to 'Diastolic_BP', 'SkinThickness' to 'Skin_Fold', 'Insulin' too 'Serum_Insulin', and 'DiabetesPedigreeFunction' to 'Diabetes_Pedigree'. These improvements were performed to improve the dataset's consistency and understandability. In the Iraqi data set feature names are renamed “AGE” has been changed to “Age”, feature extraction was done Glucose value was calculated from HbA1c.

3.2.2 Label encoding

Equations (1) and (2) transform the non-numerical qualitative values of these attributes into numerical ones so that they can be used in models. Because this work data set contains qualitative values for the Class label and Gender variables, this is essential to encode. Lable Outcome is represented as L.

$$\mathrm{gender}(x) = \begin{cases} 0, & \mathrm{if} x = \mathrm{Female} \\ 1, & \mathrm{if} x = \mathrm{Male} \end{cases} \quad (1)$$

$$L = \begin{cases} 0, \mathrm{if}\ L = N(\mathrm{NonDiabetic}) \\ 1, \mathrm{if}\ L = Y(\mathrm{Diabetic}) \end{cases} \quad (2)$$

3.2.3 K-Nearest Neighbors imputation

After Label encoding the data set has some missing elements, missing values removed means data loss occurs to maintain the original sample and perform accurate analysis to replace the missing values with some substitute. Figure 3 shows the KNN Imputation in the PIMA data set. In the dataset first row represents the feature name. 10 attributes under it have numerical and categorical values. In the KNN imputation using data imputation details are given below. Missing values are identified and filled with ‘N’ represented in green color. For example, Skinfold=2, Serum_Insulin-5. The blue color represents the categorical value of the Outcome. Example diabetic Yes-1, No-0. First, the distance between the rows is based on available elements. Euclidean distance formula used to calculate distance Dist(m,n) in equ (3).

$$\mathrm{Dist}(m, n) = \sqrt{\textstyle\sum_i (a_{mi} - a_{ni})^2} \quad (3)$$

where a_{mi} and a_{ni} are the values of feature i for instances m and n. The k value is taken as 2. After manipulating the nearest neighbor value than calculated the mean. This procedure is iteratively applied until all of the values that are not present in the dataset have been imputed. The same KNN imputation is applied to all the 3 datasets after imputed data sets are given to the ETL process.

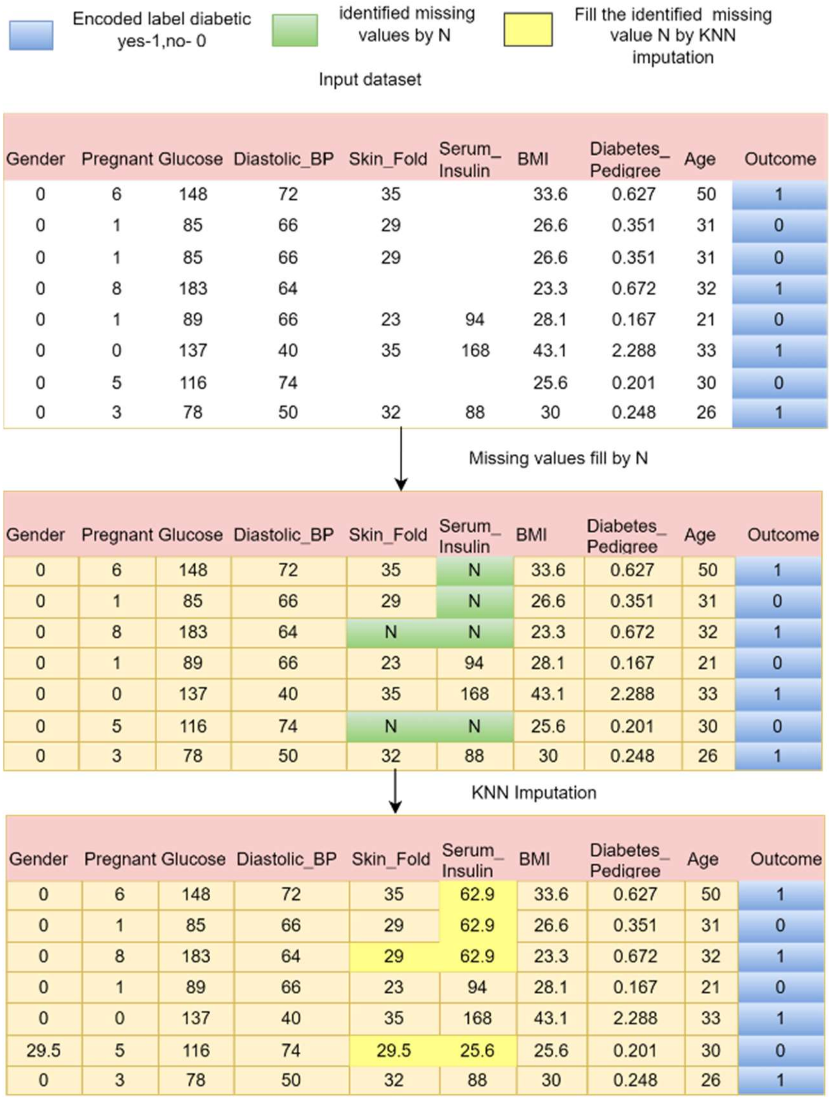
The blue color represents the encoding label of 0,1. The green color represents the missing value filled with N. The final Yellow represents the final imputation after updating the values.

3.3 Dataset Integration Using ETL

Data integration and ETL methods strive to combine and extract information from multiple sources into a single, high-quality dataset while maintaining consistency and reliability. The systematic extraction, translation, and loading of data into a single repository, which allows for advanced data analytics and rapid reporting, results in enhanced decision-making and operational efficiency. ETL (extract, transform, and load) is a popular way of transporting and transforming data from multiple sources to a destination.

A procedure with three major steps is described: Extraction retrieves data from three diverse sources of information and converts it to a single file format. Importing data into consolidated data sources is the second step in the transformation process.

The data is filtered, extracted, and verified as well as converted to different forms of data, combined from different sources, and new features are added to this preserve. In addition, the most current value in the appropriate variables is substituted for each missing value using hot deck imputation to fill in the gaps in the data. The third and last phase in the ETL method is load.


🟦 Encoded label diabetic yes-1,no- 0
🟩 identified missing values by N
🟨 Fill the identified missing value N by KNN imputation
Input dataset
Missing values fill by N
KNN Imputation

Gender	Pregnant	Glucose	Diastolic_BP	Skin_Fold	Serum_Insulin	BMI	Diabetes_Pedigree	Age	Outcome
0	6	148	72	35		33.6	0.627	50	1
0	1	85	66	29		26.6	0.351	31	0
0	1	85	66	29		26.6	0.351	31	0
0	8	183	64			23.3	0.672	32	1
0	1	89	66	23	94	28.1	0.167	21	0
0	0	137	40	35	168	43.1	2.288	33	1
0	5	116	74			25.6	0.201	30	0
0	3	78	50	32	88	30	0.248	26	1

Gender	Pregnant	Glucose	Diastolic_BP	Skin_Fold	Serum_Insulin	BMI	Diabetes_Pedigree	Age	Outcome
0	6	148	72	35	N	33.6	0.627	50	1
0	1	85	66	29	N	26.6	0.351	31	0
0	8	183	64	N	N	23.3	0.672	32	1
0	1	89	66	23	94	28.1	0.167	21	0
0	0	137	40	35	168	43.1	2.288	33	1
0	5	116	74	N	N	25.6	0.201	30	0
0	3	78	50	32	88	30	0.248	26	1

Gender	Pregnant	Glucose	Diastolic_BP	Skin_Fold	Serum_Insulin	BMI	Diabetes_Pedigree	Age	Outcome
0	6	148	72	35	62.9	33.6	0.627	50	1
0	1	85	66	29	62.9	26.6	0.351	31	0
0	8	183	64	29	62.9	23.3	0.672	32	1
0	1	89	66	23	94	28.1	0.167	21	0
0	0	137	40	35	168	43.1	2.288	33	1
29.5	5	116	74	29.5	25.6	25.6	0.201	30	0
0	3	78	50	32	88	30	0.248	26	1

Figure 3: Pima Dataset Sample Using KNN Imputation

Within this stage, the converted information is encumbered into a single dataset. During loading, preset mapping is used to create and update the final dataset. Final data is provided for the following step in the feature selection process.

3.4 Feature Selection

Finding the most pertinent features in a dataset can prove rather challenging but analyzing the attributes from varied vantages gives us a more well-rounded view. Ensemble the Fisher Score, the KS Score, and the Chi-Squared Test, the Fisher-KS-Chi2 Score (Ensemble FKCS) is a composite metric that machine learning researchers use for feature selection.

By combining the best features from each approach, this score hopes to give a reliable indicator of feature importance. Below is a detailed description of each approach and how blending them enhances our evaluation:

In Fisher Score by contrasting the variation between the classes with the variance within each class, the Fisher Score determines a feature's discriminative potential.

The overall mean (μ) of the feature is given below equ (4)

$$\mu = \frac{1}{M}\sum_{k=1}^{M} a_k \quad (4)$$

Where M is the totality numeral of sample and a_k is the value of the attribute of sample k.

The class mean μ_{cl} equ (5) and variance σ_{cl}^2 (6), for each class cl.

$$\mu_{cl} = \frac{1}{M_{cl}}\sum_{k\epsilon cl_e} a_k \quad (5)$$

$$\sigma_{cl}^2 = \frac{1}{M_{cl}}\sum_{k=CL_e}(a_k - \mu_{cl})^2 \quad (6)$$

Where M_{cl} is the quantity of sample in set cl, and CL_e is the set of indices for class CL. Between classes variance V_B is given in equ (7).

$$V_B = \sum_{cl=1}^{CL} M_{CL}(\mu_{cl} - \mu)^2 \quad (7)$$

CL is the number of classes. Within class variance V_w in equ (8).

$$V_w = \sum_{cl=1}^{CL}\sum_{k\epsilon CL_e}(a_k - \mu_{CL})^2 \quad (8)$$

$$\text{Fisher score } (FS_1) = \frac{V_B}{V_w} \quad (9)$$

Kolmogorov-Smirnov (KS) Score the largest difference between the Empirical Cumulative Distribution Functions (ECDFun) of the two classes is measured by the KS Score. ECDFun for 2 classes, for each class P and Q the value of 'a' is given in the equ (10) and (11).

$$F_P(a) = \frac{1}{M_P}\sum_{k\epsilon P} 1(a_k \leq a) \quad (10)$$

$$F_Q(a) = \frac{1}{M_Q}\sum_{k\epsilon Q} 1(a_k \leq a) \quad (11)$$

Where 1 is the indicator function, it shows the truth. M_P and M_Q represent the total number of observations in P and Q.
KS statistic represented in equ (12) $D_{P,Q}$ represent the KS score.

$$D_{P,Q} = \max_{a} | F_P(a) - F_Q(a)| \quad (12)$$

In Chi Squared Test the Observed Frequencies Construct a contingency table of observed frequencies OF_{kl} where k corresponds to the bins of the features plus l indicates the classes.
Expected Features EF_{kl} given in equ (13).

$$EF_{kl} = \frac{(\sum_l OF_{kl})(\sum_k OF_{kl})}{M} \quad (13)$$

wherever $(\sum_l OF_{kl})$the numeral of samples within the feature be bin and $(\sum_k OF_{kl})$ is the total number of samples in class l.

Chi-Squared Statistics ($\aleph^2$) represented in equ(14).

$$\aleph^2 = \sum_k\sum_l \frac{(OF_{kl} - EF_{kl})^2}{EF_{kl}} \quad (14)$$

Normalization $Nscore_k$ given in equ (15) for the sake of comparison, standardize each set of ratings to a common scale from 0 to 1.

$$Nscore_k = \frac{sc_k - mi\ (sc_k)}{\max(sc_k) - \min(sc_k)} \quad (15)$$

Where k represents the Fisher score, its score, and chi squared test. Average or weighted averaging are two aggregation methods that can be used to combine the normalized scores Comb_score given in Equ (16)

$$Comb\_score_i = \frac{nor\_F_i + nor\_KS_i + nor\_chiS_i}{3} \quad (16)$$

Where, i represent feature selection.

3.5 Optimized Classifier Selection

As part of the classifier selection is done by using Particle Swarm Optimization (PSO). It is inspired by the social behavior of birds flocking, a nature-inspired optimization algorithm. In this work, 4 classifiers are used. PSO is used to select the best classifier.

Pseudocode: The proposed PSO for classifier selection

```
Input: a set of classifiers c, p, I, w, cc, sc, k
```

Output:gbest_posi as best_classifier

```
Initialize the set of classifiers c, particles p, iterations I, Inertia weight w, cognitive coefficient cc, social coefficient sc, number of folds k
Initialize posi, velo, pbest
for each particle i in 1 to p:
    posi[i] = random(classifier)
    velo[i] = small random value
    pbest[i] = posi[i]

# Evaluate initial fitness
for each particle i in 1 to p:
    # Use Stratified K-Fold cross-validation accuracy of posi[i] to evaluate fitness
    fitness[i] = stratified_k_fold_accuracy(posi[i], k)
    pbest_fit[i] = fitness[i]

# Initialize global best
gbest_posi = posi of a particle with highest pbest_fit
gbest_fit = highest pbest_fit

for each iter t in 1 to I:
    for each particle i in 1 to p:
        # Update Velocity
        velo[i] = w * velo[i] + cc * rand() * (pbest[i] - posi[i]) + sc * rand() * (gbest_posi - posi[i])

        # Update Position
        posi[i] = posi[i] + velo[i]
        ensure posi[i] corresponds to a valid classifier in c
        # Fitness Evaluation using Stratified K-Fold cross-validation
        fit[i] = stratified_k_fold_accuracy(posi[i], k)

        # Personal pbest update
        if fit[i] >pbest_fit[i]:
            pbest[i] = posi[i]
            pbest_fit[i] = fit[i]

        # Global gbest update
        if fit[i] >gbest_fit :
            gbest_posi = posi[i]
            gbest_fit = fit[i]

return gbest_posi as best_classifier
```

3.6 Classification Method ETL-POXGB

To keep the class distribution consistent across all folds, a 5-split architecture is utilized by Stratified Cross-Validation. The model is trained on k-1 folds for each fold before being tested on the final fold. Predictions are made and a range of performance metrics, such as accuracy, precision, recall, and others, are monitored. Once all folds are completed, the mean value of each performance parameter is calculated and shown. Optimized XGBoost algorithm to train a classification model on the prepared data. This involves configuring hyperparameters, selecting features, and training the model to make accurate predictions.

4. RESULTS AND DISCUSSIONS

Model evaluation metrics are critical for determining how effectively machine learning models perform. They provide numerical metrics to help in model selection and hyperparameter adjustment. It is critical to understand which metric to use when analyzing model outcomes, as different jobs require different metrics. Some common evaluation criteria for classification tasks with discrete labels as output are: Accuracy (Accura), Precision (Precis), Recall (ReC), and F1_sco. One simple way to assess classification performance is through accuracy. It provides a rapid assessment of the model's accuracy and is derived as the ratio of correctly predicted data to total observations. The precision measures how many of all expected positive observations were found to be true. It can also be described as having a "positive predictive value".

The ratio of genuine positives to all expected positives is referred to as recall or sensitivity. These strategies are particularly effective in unbalanced datasets. The F1 Score is calculated by smoothly averaging recall and precision. It strikes a nice balance between the two measures and is useful for dealing with both false positives and false negatives. Evaluation metrics are critical for determining the effectiveness of machine learning models. Numerical metrics can help with model selection and hyperparameter tuning. Because different tasks necessitate different metrics, it is vital to understand which metric to employ when assessing model performance.

To ensure the validity of our diabetic prediction model, addressed potential threats to validity and carefully chose critique criteria for a thorough examination. Internal validity concerns such as preprocessing bias and overfitting were reduced by rigorous validation approaches, while external validity threats such as dataset generalizability and class imbalance were managed through the integration of varied datasets and the use of multiple evaluation metrics. Construct validity was guaranteed by selecting clinically relevant characteristics and conducting a balanced assessment utilizing comprehensive performance metrics (Precision, Recall, F1-score, ROC AUC, and MCCoeff). For a fair comparison, the model was compared with proven techniques, revealing improved predicted performance and dependability.

The Accura, Precis, ReC, F1_sco, and ROC AUC can be quantitatively expressed using the equations (17) through (21) below.

$$\text{Accura} = \frac{\text{TrPs+TrNg}}{\text{TrPs + TrNg + FlPs + FlNg}} \quad (17)$$

$$\text{Precision(Precis)} = \frac{\text{TrPs}}{\text{TrPs +FlP}} \quad (18)$$

$$\text{Recall(ReC)} = \frac{\text{TrPs}}{\text{TrPs +FlN}} \quad (19)$$

$$\text{F1\_sco} = 2 * \frac{\text{Precis}*\text{ReC}}{\text{Precis +ReC}} \quad (20)$$

$$\text{ROCAUC} = \int_0^1 \text{TrPsR(FlPsR)d(FlPsR)} \quad (21)$$

$$\text{Cohen'sKappa(CK)} = \frac{P_o-P_e}{1-P_e} \quad (22)$$

In equ (22) shows the CK value the P_o represents the observed prediction of the actual label, and P_e represents the expected agreement chance.

$$\text{MCCoeff} = \frac{\text{TrPs} * \text{TrNg} - \text{FlPs} * \text{FlNg}}{\sqrt{(\text{TrPs + FlPs})(\text{TrPs + FlNg})(\text{TrNg + FlPs})(\text{TrNg + FlNg})}}$$

(23)

MCCoeff is in equ (23) it shows a value that represents the connection between the experimental and predicted double classifications.

Table 1: Performance Comparison of All Machine Learning Models on The Three Integrated Dataset.

Classifiers	***Accura (%)***	***Precis (%)***	***ReC (%)***	***F1_sco (%)***	***ROC AUC(%)***	***CK (%)***	***MCCoeff (%)***
Logistic Regression (LReg)	0.7017	0.5432	0.8602	0.6659	0.8119	0.4203	0.4569
Random Forest (RFo)	0.9188	0.8537	0.9232	0.8871	0.9749	0.8238	0.8254
XGBoost (XGBo)	0.9363	0.8831	0.9401	0.9107	0.9692	0.8613	0.8623
AdaBoost (ABo)	0.7596	0.6537	0.6467	0.6502	0.8228	0.4670	0.4670
Bagging (Bagg)	0.9135	0.8789	0.8694	0.8741	0.9702	0.8082	0.8083
Gradient Boosting (GBo)	0.8227	0.7427	0.7450	0.7439	0.8929	0.6083	0.6083
SVM	0.7691	0.6343	0.7834	0.7010	0.8363	0.5163	0.5240
K-Nearest Neighbors (K-NN)	0.7803	0.7075	0.6206	0.6612	0.8392	0.4996	0.5020
Decision Tree (DeT)	0.9018	0.8795	0.8295	0.8538	0.8847	0.7799	0.7807
Naive Bayes (NBa)	0.6847	0.5836	0.3057	0.4012	0.7246	0.2146	0.2353
Proposed Model ETL-POXGB	0.9716	0.9731	0.9691	0.971	0.9929	0.9431	0.9433

4.1. Performance Analysis of the Proposed ETL-POXGB Model

Table 1 enumerates the Integrated three-dataset performance analysis of the entire model and also the proposed model ETL-POXGB. The models RFo, XGBo, Bagg, and DeT achieved the accura of 90% above comparing the proposed ETL-POXGB model which gave a better accuracy of 97.1%. Figure 4 represents the pictorial representation of Table 1. The ETL-POXGB model outperforms other classifiers in accuracy, precision, recall, F1-score, and robustness. It achieves high accuracy with an F1-score of 0.971, maintains a balance between precision and recall, and has superior discriminatory power. Its robustness and agreement with actual labels make it a reliable and efficient approach for diabetic risk prediction.

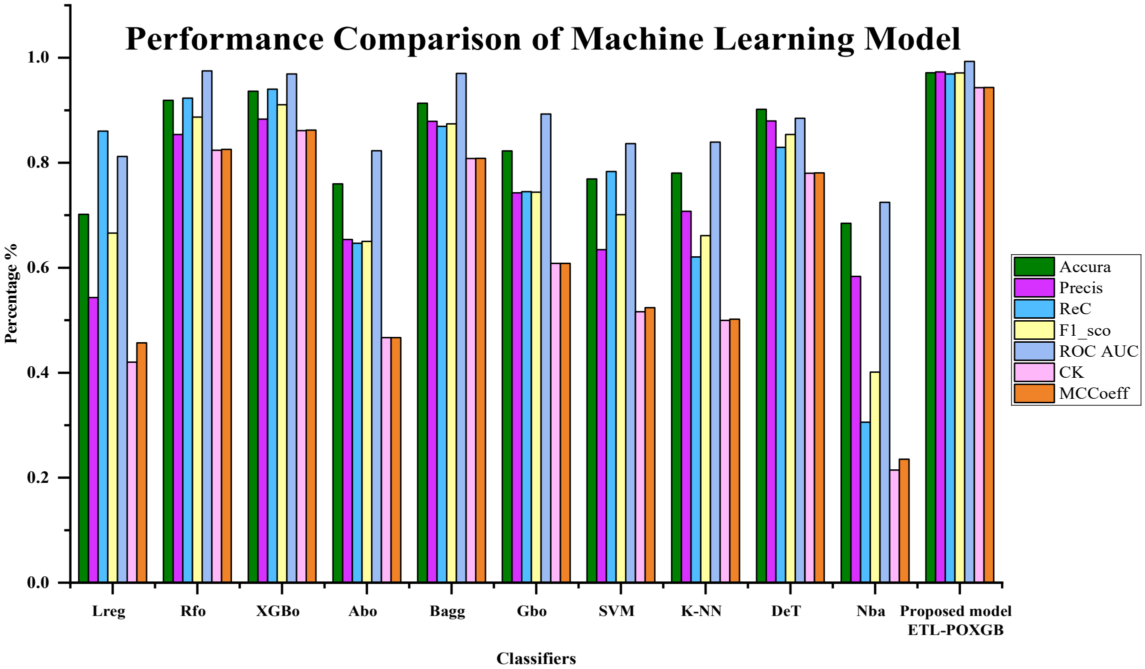


Figure 4: Performance Comparison of Machine Learning Model

Table 2: Stratified KFold Using Performance Comparison of All Models on The Three Integrated Dataset.

Classifiers	***Accura (%)***	***Precis (%)***	***ReC (%)***	***F1_sco (%)***	***ROC AUC (%)***	***CK (%)***	***MCCoeff (%)***
Logistic Regression (LReg)	0.7850	0.7940	0.7593	0.7762	0.8644	0.5696	0.5702
Random Forest (RFo)	0.9536	0.9525	0.9529	0.9527	0.9904	0.9071	0.9071
XGBoost (XGBo)	0.9613	0.9641	0.9567	0.9604	0.9853	0.9225	0.9226
AdaBoost (ABo)	0.7980	0.8003	0.7847	0.7923	0.8879	0.5958	0.5961
Bagging (Bagg)	0.9504	0.9597	0.9384	0.9489	0.9854	0.9007	0.9010
Gradient Boosting (GBo)	0.8455	0.8475	0.8372	0.8420	0.9260	0.6909	0.6914
SVM	0.8137	0.8128	0.8069	0.8096	0.8884	0.6272	0.6276
K-Nearest Neighbors (K-NN)	0.8392	0.8385	0.8329	0.8356	0.9233	0.6782	0.6784
Decision Tree (DeT)	0.9509	0.9584	0.9411	0.9495	0.9507	0.9017	0.9021
Naive Bayes (NBa)	0.7747	0.7911	0.7350	0.7618	0.8403	0.5486	0.5502
Proposed Model ETL-POXGB	0.9716	0.9731	0.9691	0.9710	0.9929	0.9431	0.9433

Table 2 depicts the Integrated 3 dataset performance analysis among the Stratified KFold cross-validation maintaining the consistency of all the model performances. All the model's accuracy is increased compared the Table 1. Figure 5 shows the all-model performance with the Stratified KFold cross-validation chart. Table 3 highlights by comparison of various machine learning models that were compared to the 3 Integrated Datasets, with ETL-POXGB being identified as the top performer with high accuracy and precision. Strong classification capabilities were also demonstrated by XGBoost and Random Forest. An accuracy of 0.9716, a precision of 0.9731, then a CK score of 0.9431 was attained by ETL-POXGB.

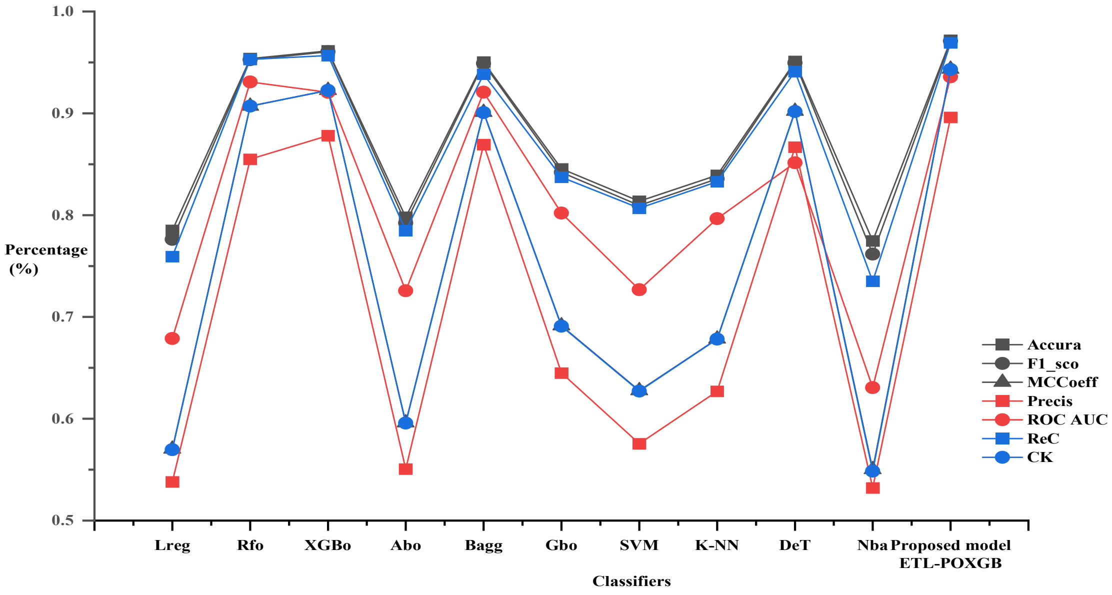


Figure 5: Performance Comparison of Stratified KFold Cross Validation with All Machine Learning

Table 3: Performance Comparison of FS-KS-Chi2 Ensem and Stratified KFold with All Machine Learning Models on The Three Integrated Dataset.

Classifier	***Accura (%)***	***Precis (%)***	***ReC (%)***	***F1_sco (%)***	***ROC AUC (%)***	***CK (%)***	***MCCoeff (%)***
Logistic Regression (LReg)	0.7893	0.7971	0.7658	0.7811	0.8716	0.5781	0.5787
Random Forest (RFo)	0.9602	0.9576	0.9616	0.9595	0.9941	0.9204	0.9204
XGBoost (XGBo)	0.9684	0.9692	0.9665	0.9678	0.9933	0.9368	0.9369
AdaBoost (ABo)	0.8182	0.8215	0.8058	0.8133	0.9117	0.6362	0.6366
Bagging (Bagg)	0.9549	0.9644	0.9432	0.9536	0.9919	0.9097	0.9101
Gradient Boosting (GBo)	0.8774	0.8866	0.8616	0.8735	0.9495	0.7546	0.7555
SVM	0.8548	0.8691	0.8307	0.8489	0.9235	0.7093	0.7109
K-Nearest Neighbors (K-NN)	0.8280	0.8425	0.7999	0.8204	0.9230	0.6556	0.6568
Decision Tree (DeT)	0.9528	0.9625	0.9405	0.9513	0.9525	0.9054	0.9059
Naive Bayes (NBa)	0.7399	0.8152	0.6079	0.6963	0.8377	0.4772	0.4930
Proposed Model ETL-POXGB	0.9716	0.9731	0.9691	0.9710	0.9929	0.9431	0.9433

XGBoost and Random Forest were also found to be highly effective in delivering accurate and reliable results. The entire model Stratified KFold cross-validation is applied.

4.2 Comparative Analysis of Proposed ETL-POXGB Over Existing Machine Learning Models

The proposed work includes K-Nearest Neighbours imputation to reduce bias and improve data quality in three datasets. Further, combined diverse datasets using ETL, performed hot deck imputation, and used ensemble methods such as the Fisher score, Kolmogorov-Smirnov score, and Chi-Square for diabetic feature selection. Finally, implemented PSOptimization for optimal prediction of the disease. This work evaluated the effectiveness of our diabetes prediction system by comparing it to another state of artworks in terms of the performance measures that provide a comprehensive understanding of the model's performance.

The ETL-POXGB model has been evaluated and compared with recent literature, achieving an accuracy of 97.16%, outperforming AHDHS-Stacking (93.09%), DP-UCE (87.41%), and RF (91%). It also shows higher precision (97.31%) and recall (96.91%), ensuring better sensitivity and specificity than XGBoost with ADASYN (88.5%). With the highest MCC of 94.33%, it demonstrates strong agreement with actual labels. The model's improved generalization across various datasets highlights its effectiveness in diabetic risk prediction and medical diagnosis.

Table 4: Comparison of Proposed Model Performance with Existing Similar Prediction Work

Ref	*Method*	*Dataset*	*Accura (%)*	*Precis (%)*	*ReC (%)*	*F1_sco (%)*	*MCCoeff (%)*
[21]	AHDHS-Stacking	PID, CWMD	93.09	93.22	91.6	92.25	84.79
[22]	DP-UCE	PID	87.41	86.47	87.25	87.69	87.65
[23]	ANN-GA	PID	86.48	82.38	85.48	85.69	86.85
[24]	AIDM	PIMA, LMCH	84.47	80.45	81.29	83.48	82.34
[25]	STACK-GRAPE	WMH, PID	80.3	78.9	68.2	72.9	78.1
[26]	RF	PID	89.86	89.18	89.77	89.91	93.72
[27]	DT+RF+XGB+LGB	PIMA	96.89	97.81	99.23	98.87	-
[28]	XGBoost with ADASYN	PIMA, LTF	88.5	82	80	81	-
[29]	RF	PIMA	91	-	-	-	-
[30]	MLP and LSTM are fine-tuned	PIMA	86.083	86.6	85.1	-	-
Proposed Model ETL-POXGB		PIMA, Iraqi, Frankfurt	97.16	97.31	96.91	97.1	94.33

Table 4 represents the comparison of the model performance with existing prediction work. It represents the different models and the different datasets used. The proposed model (ETL-POXGB) outperforms previous models in accuracy (97.16%), recall (96.91%), and precision (97.31%) while maintaining a high MCC score (94.33%).

The Proposed model achieved the highest accuracy, surpassing the existing methods, DT+RF+XGB+LGB, AHDHS-Stacking, and RF with their accuracy of 96.89%, 93.09%, and 89.86% respectively. Thus, increased accuracy demonstrates the robustness of the ETL-POXGB framework in effectively handling diverse datasets. Similarly, the higher precision rate helps to reduce the false positives, therefore making the system highly reliable for prediction. ETL-POXGB integrates advanced preprocessing techniques, improving data quality and feature selection with metaheuristic optimizations leading to improved performance. These findings emphasize its efficacy in diabetic prediction, making it a more dependable and generalizable model for real-world use.

5. CONCLUSION

The development of accurate and early prediction models is essential for the most effective disease management, opportune therapies, and the lowest possible health risks, as the global diabetes burden continues to increase. This work contributes to this goal by proposing a model that enhances forecast accuracy while maintaining dependability across multiple datasets. This research work compared the proposed model ETL-POXGB, with the existing eleven machine learning models, including LReg, RFo, XGBo, ABo, Bagg, GBo, K-NN, DeT, and NBa in diabetics prediction. Three different data sets were imputed using KNN imputation and then combined by utilizing the ETL procedure. The selected features are subsequently added to the optimal classifier model. According to the test results, ETL-POXGB achieved a good accuracy rate of 97.16%. Furthermore, the system excelled in other evaluation criteria such as Precis, ReC, F1_sco, and MCCoeff. The analysis of feature importance showcased how independent features influenced the final forecast.

Even though the proposed approach extracts the relevant features effectively for the dataset used in this work it may be a challenge to produce high accuracy for the higher dimensional dataset. Therefore, to address this challenge, future work can be extended to integrate deep learning techniques to handle feature extraction more efficiently. In addition, exploring advanced evolutionary-based optimization for classification can also be considered as a potential suggestion for improvement. The proposed method could be applied to other medical datasets with similar characteristics to broaden the scope of this research.

REFERENCES:

[1] Felizardo, V., Garcia, N.M., Megdiche, I., Pombo, N., Sousa, M. and Babič, F., 2023. Hypoglycaemia prediction using information fusion and classifiers consensus. *Engineering Applications of Artificial Intelligence*, *123*, p.106194.

[2] Bhat, S.S., Banu, M., Ansari, G.A. and Selvam, V., 2023. A risk assessment and prediction framework for diabetes mellitus using machine learning algorithms. *Healthcare Analytics*, p.100273.

[3] Alhakeem, Z.M., Hakim, H., Hasan, O.A., Laghari, A.A., Jumani, A.K. and Jasm, M.N., 2023. Prediction of diabetic patients in Iraq using binary dragonfly algorithm with long-short term memory neural network. *AIMS Electronics & Electrical Engineering*, *7*(3).

[4] El Massari, H., Sabouri, Z., Mhammedi, S. and Gherabi, N., 2022. Diabetes prediction using machine learning algorithms and ontology. *Journal of ICT Standardization*, *10*(2), pp.319-337.

[5] Patil, R., Tamane, S., Rawandale, S.A. and Patil, K., 2022. A modified mayfly-SVM approach for early detection of type 2 diabetes mellitus. *Int. J. Electr. Comput. Eng*, *12*(1), pp.524-33.

[6] Mahesh, T.R., Kumar, D., Kumar, V.V., Asghar, J., Bazezew, B.M., Natarajan, R. and Vivek, V., 2022. Blended ensemble learning prediction model for strengthening diagnosis and treatment of chronic diabetes disease. *Computational Intelligence and Neuroscience*, *2022*.

[7] Ali, M.S., Islam, M.K., Das, A.A., Duranta, D.U.S., Haque, M. and Rahman, M.H., 2023. A novel approach for best parameters selection and feature engineering to analyze and detect diabetes: Machine learning insights. *BioMed Research International*, *2023*.

[8] Beghriche, T., Djerioui, M., Brik, Y., Attallah, B. and Belhaouari, S.B., 2021. An efficient prediction system for diabetes disease based on deep neural network. *Complexity*, *2021*, pp.1-14.

[9] Patra, R., 2021, February. Analysis and prediction of Pima Indian Diabetes Dataset using SDKNN classifier technique. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1070, No. 1, p. 012059). IOP Publishing.

[10] Ganie, S.M., Pramanik, P.K.D., Bashir Malik, M., Mallik, S. and Qin, H., 2023. An ensemble learning approach for diabetes prediction using boosting techniques. *Frontiers in Genetics*, *14*, p.1252159.

[11] Tasin, I., Nabil, T.U., Islam, S. and Khan, R., 2023. Diabetes prediction using machine learning and explainable AI techniques. *Healthcare technology letters*, *10*(1-2), pp.1-10.

[12] Yuan, Z., Ding, H., Chao, G., Song, M., Wang, L., Ding, W. and Chu, D., 2023. A Diabetes Prediction System Based on Incomplete Fused Data Sources. *Machine Learning and Knowledge Extraction*, *5*(2), pp.384-399.

[13] Bassam, G., Rouai, A., Ahmad, R. and Khan, M.A., 2023. Diabetes Prediction Empowered with Multi-level Data Fusion and Machine Learning. *International Journal of Advanced Computer Science and Applications*, *14*(10).

[14] Sravanthi, A.L., Al-Ashmawy, S., Kaur, C., Al Ansari, M.S., Saravanan, K.A. and Vuyyuru, V.A., 2023. Utilizing Multimodal Medical Data and a Hybrid Optimization Model to Improve Diabetes Prediction. *diabetes*, *14*(11).

[15] Kulkarni, M. and Deore, S., 2024. Predicting Microvascular Complications in Diabetic Mellitus Using Improved Enhanced Coati Optimizer. *International Journal of Computing and Digital Systems*, *16*(1), pp.1-18.

[16] Ahmed, U., Issa, G.F., Khan, M.A., Aftab, S., Khan, M.F., Said, R.A., Ghazal, T.M. and Ahmad, M., 2022. Prediction of diabetes

empowered with fused machine learning. *IEEE Access*, *10*, pp.8529-8538.

[17] Reddy, S. and Mahesh, G., 2021. Risk assessment of type 2 diabetes mellitus prediction using an improved combination of NELM-PSO. *EAI Endorsed Transactions on Scalable Information Systems*, *8*(32).

[18] Wee, B.F., Sivakumar, S., Lim, K.H., Wong, W.K. and Juwono, F.H., 2023. Diabetes detection based on machine learning and deep learning approaches. *Multimedia Tools and Applications*, pp.1-33.

[19] Zhou, H., Xin, Y. and Li, S., 2023. A diabetes prediction model based on Boruta feature selection and ensemble learning. *BMC bioinformatics*, *24*(1), p.224.

[20] Abnoosian, K., Farnoosh, R. and Behzadi, M.H., 2023. Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. *BMC bioinformatics*, *24*(1), p.337.

[21] Zhang, Z., Lu, Y., Ye, M., Huang, W., Jin, L., Zhang, G., Ge, Y., Baghban, A., Zhang, Q., Wang, H. and Zhu, W., 2024. A novel evolutionary ensemble prediction model using harmony search and stacking for diabetes diagnosis. *Journal of King Saud University-Computer and Information Sciences*, *36*(1), p.101873.

[22] Prabhakar, G., Chintala, V.R., Reddy, T. and Ruchitha, T., 2024. User-cloud-based ensemble framework for type-2 diabetes prediction with diet plan suggestion. *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, *7*, p.100423.

[23] Rajagopal, A., Jha, S., Alagarsamy, R., Quek, S.G. and Selvachandran, G., 2022. A novel hybrid machine learning framework for the prediction of diabetes with context-customized regularization and prediction procedures. *Mathematics and Computers in Simulation*, *198*, pp.388-406.

[24] Olisah, C.C., Smith, L. and Smith, M., 2022. Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Computer Methods and Programs in Biomedicine*, *220*, p.106773.

[25] Kumar, A. and Kaur, K., 2024. A Novel MCDM-Based Framework to Recommend Machine Learning Techniques for Diabetes Prediction. *International Journal of Engineering & Technology Innovation*, *14*(1).

[26] Salih, M.S., 2024. Diabetic Prediction based on Machine Learning Using PIMA Indian Dataset. *Communications on Applied Nonlinear Analysis*, *31*(5s), pp.138-156.

[27] Talari, P., N, B., Kaur, G., Alshahrani, H., Al Reshan, M.S., Sulaiman, A. and Shaikh, A., 2024. Hybrid feature selection and classification technique for early prediction and severity of diabetes type 2. *Plos one*, *19*(1), p.e0292100.

[28] Tasin, I., Nabil, T.U., Islam, S. and Khan, R., 2023. Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters*, *10*(1-2), pp.1-10.

[29] Usha, V. and Rajalakshmi, N.R., 2023, September. Insights into Diabetes Prediction: A Multi-Algorithm Machine Learning Analysis. In *2023 4th International Conference on Smart Electronics and Communication (ICOSEC)* (pp. 1207-1212). IEEE.

[30]Butt, U.M., Letchmunan, S., Ali, M., Hassan, F.H., Baqir, A. and Sherazi, H.H.R., 2021. Machine learning based diabetes classification and prediction for healthcare applications. *Journal of healthcare engineering*, *2021*(1), p.9930985.