International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 16 (2024) pp. 184–196

Article

The Effect of Anomaly Detection and Data Balancing in Prediction of Diabetes

Aruna Devi B. * and Karthik N.

Department of Computer Science and Engineering, National Institute of Technology Puducherry, Karaikal 609609, India; nkarthikapce@gmail.com
* Correspondence author: arunaakshal@gmail.com

Received date: 5 March 2024; Accepted date: 21 March 2024; Published online: 24 May 2024

Abstract: Diabetes is a health condition spurred by elevated blood glucose, commonly called blood sugar. In developing countries, diabetes is the most common illness. Expert medical intervention and prompt diagnostics are crucial measures in mitigating the effects of diabetes. The proliferation of databases in the healthcare industry presents numerous opportunities for artificial intelligence and machine learning technologies. Despite the availability of numerous medical devices, medical errors remain a major problem in the healthcare industry. Medical data with anomalous values can lead to wrong decisions. Anomaly detection is frequently employed in datasets to locate and eliminate anomalies. On the other hand, identifying and evaluating the outlier pattern may enhance a learning algorithm's medical decisions and precision. This paper presents a novel strategy for diabetes prediction based on KNN imputation, Hybrid Sampling and Anomaly Detection. This model increases the detection rate of diabetes in the Pima Indian diabetes dataset. This work utilized five unsupervised anomaly detection algorithms and five supervised machine learning algorithms to perform diabetes prediction. This work was assessed under four conditions: diabetes prediction without anomaly detection in the unbalanced dataset, with a balanced dataset and without anomaly detection, with anomaly detection in the unbalanced dataset, and with anomaly detection in the balanced dataset. Results confirmed that the Isolation Forest and Random Forest outperform the other machine learning models in diabetes prediction with 99.23% accuracy and a precision of 0.99. The findings demonstrated that all compared methods could detect anomalous data and produce consistent outcomes across the different algorithms. The results of our experiments show that our method works better at identifying anomalies and highlighting the significance of dataset balancing and anomaly detection in diabetes prediction.

Keywords: KNN Imputation; anomaly detection; hybrid sampling; diabetes prediction; supervised learning; unsupervised learning

1. Introduction

A vital initial phase in treating a disease is diagnosing it. Specifically, it is essential to diagnose diseases like diabetes as early as possible. This condition occurs when the body produces insufficient insulin or manages it ineffectively. Significantly, failing to detect diabetes at an early age not only affects the prognosis for the illness but also creates the conditions necessary for the emergence of additional chronic illnesses like kidney disease [1]. 8.5% of people under the age of 18 had diabetes in 2014. Around 1.6 million people died from diabetes worldwide in 2016, 2012 saw 2.2 million deaths worldwide due to diabetes. Diabetes had claimed more lives in 2019 than it had in previous years [2]. Diabetes is linked to a wide range of risk factors, including smoking habits, aging, family history, being overweight, eating poorly, and not getting enough exercise [3]. Early detection of these chronic illnesses has two advantages: it helps to prevent future medical expenses, and, at the same time, it lessens the chance of worsening health complications, which guarantees the preservation of a patient's quality of life [4]. Health sciences, computer science and information science are applied in health informatics to handle and deliver data for clinical practice. Additionally, Medical data analysis tools help medical professionals acquire, exchange

and apply data and ICT expertise more quickly to enhance the quality of care [5]. Using traditional methods to analyze and manage the vast array of multi-source data and information spawned daily by healthcare providers presents enormous challenges [5]. The correct analysis of this data to extract valuable insights is aided by machine learning (ML) techniques. Prognostication of disease, diagnosis, therapy, and clinical workflow are the four main areas of healthcare where ML can drive improvement. However, the trustworthiness of the dataset used to train and assess the models significantly impacts the performance of machine learning models. Thus, it is crucial to comprehend the various facets of dataset quality in machine learning to guarantee the successful implementation of ML projects. Finding and addressing quality issues with the dataset, such as missing, anomalous, and balanced data, is a critical component of data quality [6].

In this work, we explore the factors contributing to the quality of the Pima Indian diabetes dataset. We examine the importance of anomaly or outlier detection and balancing datasets and how they impact the final prediction of diabetes made by ML models. A data point that significantly deviates from the norm or from other observations is called an outlier. Anomalies are generally caused by outside variables, like malfunctioning sensors or outside attacks. A detection algorithm's job is locating the anomaly and categorizing or deducing its cause [7]. A dataset is considered a class imbalance if the distribution of its majority and minority classes is unequal. When learning from unbalanced datasets, there is a bias in favor of the prevailing classes, whose labeled samples are more plentiful than those of the underrepresented minority class. Uneven class distribution and the data's inherent properties reduce the classifier's performance [8,9]. Ground truth labels on the data determine the best machine-learning solution for a given issue. Supervised approaches are appropriate when the data are labeled, while unsupervised techniques enable the analysis of unlabeled data [10]. Copious machine learning models can be used to balance datasets and detect anomalies. The process involved in this work is summarized as follows:

- Utilizing a machine learning approach to predict diabetes through meticulous preprocessing of the dataset. The approach was assessed under four conditions. (1) With unbalanced dataset and without anomaly detection, (2) With balancing dataset and without anomaly detection (3) Without balancing dataset and with anomaly detection, and (4) With anomaly detection and with balancing dataset. KNN imputation was utilized to replace the missing values within the dataset.
- This work focuses on five unsupervised algorithms for anomaly detection, namely Local outlier factor (LOF), Isolation Forest (iForest), One-Class SVM (OCSVM), Elliptical envelope (eEnvelope), and Lightweight On-line Detector of Anomalies (LODA).
- To address the issue of data imbalance, hybrid sampling was utilized, namely Synthetic Minority Oversampling Technique and Edited Nearest Neighbor (SMOTE-ENN), and for making diabetes predictions five supervised ML algorithms called K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF), eXtreme Gradient Boosting (XGB) and RUSBoost.
- Accuracy, precision, recall, and F1-score are the measures used to assess the performance of the proposed approach.

This is the structure of the remaining portion of the paper. The Literature review is presented in Section II. A detailed exposition of the suggested strategy is provided in Section III. A framework for the suggested prediction model is outlined, and several unsupervised anomaly detection algorithms, SMOTE-ENN, and supervised machine learning models are examined. The analysis and discussion of the experiment are presented in Section IV. It offers the dataset's statistical description along with comparable performance metrics. Section V concludes the paper and outlines a few current research trends in the relevant field.

2. Literature Survey

This section discusses the various ML models for anomaly detection, balancing datasets, and diabetes prediction. Hanaa Torkey et al. [11] integrated RF mean, class mean for imputation, interquartile range (IQR) for anomaly detection, and deep learning for outlier repair to create a tidy and comprehensive dataset. They used the Pima Indian Diabetes dataset to evaluate their model. The classifiers used are RF, SVM, Decision Tree (DT), and Naïve Bayes (NB). Assessment criteria like accuracy, Root Mean Square Error (RMSE), F1 Score, and Area Under Curve (AUC) were used. The accuracy of the proposed outlier repair was 97.41% by RF. This work did not focus on the imbalance of the dataset. Samariya et al. [12] identified outlying features, this work seeks to both efficiently and effectively identify anomalies and explain their classification as anomalies. namely LOF, iForest, Sp, and isolation using Nearest Neighbor Ensemble are the anomaly detection methods used, and outlying aspect mining algorithms. Finally, they assessed their effectiveness using sixteen real-world healthcare datasets, with Pima Indian Diabetes being one among them. SiNNE, the most recent isolation-based outlying aspect mining measure, performs exceptionally well. Edin Šabić et al. [9] proposed a method to investigate the ability of two are

unsupervised and three supervised algorithms to identify abnormalities in heart rate data. Then, using actual heart rate data, these algorithms were assessed. Results showed that both the random forests and the local outlier factor algorithms performed well in spotting deviations in heart rate, with each model exhibiting good generalization from training on synthetic heart rate data to actual heart rate data. Class imbalance learning is a pertinent research topic in machine learning and data mining. Medical and healthcare-related datasets are inherently unbalanced. Neglecting the imbalance mitigates the classifier's effectiveness which hinders the identification of unusual instances in crucial domains like disease screening, analyzing severity, identifying adverse drug reactions, classifying cancer malignancy, and identifying uncommon chronic illnesses in the general population.

Mirza Muntasir Nishat et al. [13] investigate the survivability of heart failure using six supervised ML algorithms. Several classifiers are used: DT, LR, RF, KNN, SVM, and Gaussian Naive Bayes (GNB). The synthetic minority oversampling technique and edited nearest neighbor (SMOTE-ENN) data resampling technique are utilized. With a test accuracy of 90%, the research outcomes unequivocally show that Random Forest Classifier (RFC), when combined with SMOTE-ENN and conventional scaling technique, outperforms all other methods. Oladimeji et al. [14] suggested an integrated machine learning-based strategy for predicting heart failure patients' chances of survival. In order to address the class imbalance in the classification dataset, the integrated approach ranks significant predictive features. The Random Forest algorithm showed 83.18% accuracy, which was the highest. The Pima Indian diabetes dataset and a private dataset were utilized in [15]. In this work, the mutual information feature selection algorithm has been used. A model that employs extreme gradient boosting a semi-supervised model has predicted the insulin features found in the private dataset. The issue of class imbalance has been addressed using SMOTE and ADASYN techniques. The authors deployed a range of ensemble techniques and machine learning classification methods, including DT, SVM, RF, KNN, and LS, to ascertain which algorithm yields the most accurate predictions. With an accuracy of 81%, the proposed system yielded the best results in the XGB classifier using the ADASYN approach. Wang et al. [16] used the oversampling technique to balance the data class distribution and classify diabetes mellitus. This work uses random forest (RF) classifiers to generate predictions, while the Naïve Bayes algorithm is used as a substitute for the absence of values. On the PID dataset, the proposed work yields an accuracy rate of 87.10%. Chandrashekhar Azad et al. [17] recommended a prediction model for PMSGD with four layers. Preprocessing, which comprises the first layer, includes handling missing values, detecting outliers, and minority class up-sampling. Correlation and genetic algorithms are used for feature selection in the second layer. In the third layer the proposed model is trained, and the fourth layer evaluates its effectiveness. An accuracy rate of 82.12% is obtained by the suggested algorithm.

Using the Pima Indian dataset, Ramesh et al. [18] developed a remote and automated diabetes prognosis tool. The authors used a variety of parameter scaling, feature choice, and SMOTE preliminary data processing techniques. The maximum accuracy that SVM with RBF kernel could achieve was 83.2%. A proposed machine learning framework is used in an Android app. NonsoNnamoko et al. [19] proposed an approach for data preprocessing by embedding the knowledge of outliers and SMOTE. IQR was used for outlier detection and the outliers were replaced using synthetic data generated by SMOTE. This work showed the accuracy of 89.5% and F1 score of 0.895. Fitriyani et al. [20] developed a disease prognosis framework. The suggested model combines an ensemble approach, the synthetic minority oversampling technique tomek link (SMOTETomek), and isolation forest (iForest) outlier detection. The outlier detection is eliminated using the iForest technique. The ensemble approach is taken into account when making predictions. Decision trees (DT), support vector machines (SVM), and multilayer perceptrons (MLP) are used as first-level classifiers in this work. As a second layer classifier, LR classifiers operate in the meantime. Anand Kumar Srivastava et al. [21] suggested a blended diabetes prediction framework addressing missing values using K-Mean++ based data imputation technique and outliers' detection using ABC based outlier detection and results were determined using LS-SVM classifiers. Evaluation measures like AUC, sensitivity, specificity, accuracy and kappa were used. Most of the real-world dataset are imbalanced and contains anomalies.

Uma R. Godase et al. [22] shows how different datasets' classification performance is affected by a higher degree of class imbalance and put forward the classification strategy that combines a diverse classifier ensemble (CE) with the data level technique. Vibhuti Sharma [23] have put forth that a number of machine learning methods, such as DeepRisk and transfer learning, but the majority of research focuses on supervised learning methods. LR, RF, and SVM are the most popular models used as baselines and the evaluation metrics such as Root Mean Square Error (RMSE), F1 measure, Area under the Curve (AUC), Accuracy, Precision and Recall are mostly used. Rathod S R et al [24] proposed a model that recognises heart disease using he heart rate variability (HRV) parameters, four machine learning algorithms such as naïve bayes, KNN, Artificial Neural Network (ANN) and SVM were trained in which KNN produced 0.94 accuracy.

Dataset Annthyroid has 7.4% of anomalies, Arrhythmia dataset contains 14% of anomalies, BreastW has 34% of anomalies, 8.32% of anomalies were present in Cardiotocography dataset, 34% of anomalies in diabetes dataset, dataset Hepatitis has 16.25 of anomalies, Vertebral dataset has 12% of anomalies, 7.5% anomalies were present in WBC dataset and WPBC has 23.73% of anomalies [12]. According to [25] Pima Indian diabetes dataset is imbalanced with 268 positive class and 500 negative class, WPBC dataset contains 45 positive class and 149 negative class, WDBC dataset contains 212 positive class and 357 negative class and Breast-cancer-Wisconsin contains 239 positive class and 444 negative class.

Through the observation from the works mentioned above, we have found that medical datasets frequently experience problems with missing values, identifying outliers, and class imbalance, all of which can impact the classification system's performance. In most healthcare predictions, precise preprocessing is not carried out, leading to a decline in model performance, unreliable models, and misleading conclusions. In this work, we aim to predict diabetes by addressing the data quality issues in the dataset, increasing the model's performance. We also aim to examine the importance of outlier detection and balancing data by performing predictions under four conditions. To the extent of our current literature survey, this is the work done for predicting diabetes under four different conditions based on dataset quality.

3. Proposed Work

This work focuses on devising an approach for accurately predicting diabetes by precisely preprocessing the dataset. A subpar dataset can lead to misprediction and wrong decisions; hence, it is essential to preprocess the medical dataset. In this approach, five unsupervised anomaly detection algorithms were utilized to detect and remove anomalous data, and five supervised learning algorithms were used to predict diabetes. The visual representation of the proposed approach is presented in Figure 1. The proposed strategy has the following steps: (1) KNN imputation, (2) Anomaly detection and removal, (3) Hybrid Sampling, (4) Diabetes Prediction, (5) Evaluation. We go over each step of the suggested approach in the following sections. The architecture of the proposed approach is given in Figure 2.

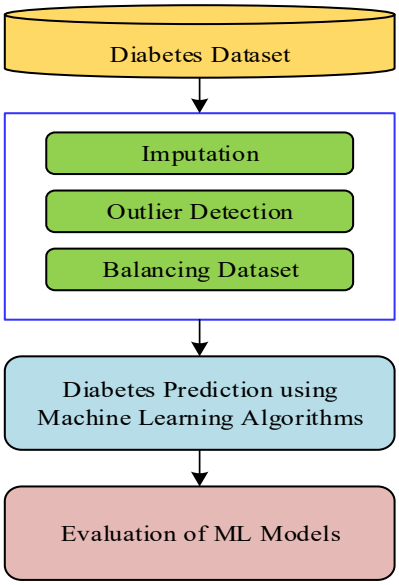

Figure 1. Block diagram of proposed work.

3.1. KNN Imputation

A frequent issue that can render the prediction system inefficient is the absence of data. Missing or zero values has to be replaced to make the prediction model effective. In this work, the dataset employed to train our model has zero values in various attributes. A publicly available dataset is used and the depiction of the dataset is given in Section IV. The model's ability to perform can be greatly affected by the method chosen to fill in these missing values, so choosing cautiously is crucial [26,27]. KNN imputation algorithm was used to address this issue. The steps involved in imputing the null values are.

Step 1: Find the zero values in the dataset and replace them with NaNs.

Step 2: Using elbow method find the K which denotes the number of nearest neighbors to be considered while imputing the missing values. In this work k value was taken as 17.

Step 3: Distance is calculated between missing values and all other data points. The distance function used here is Euclidean distance. The Euclidean distance can be found by increasing the weight of the coordinates that are not missing and disregarding the missing values. Euclidean distance can be calculated by using the following Equation (1).

$$D_{xy=}\sqrt{Weight * \ squared\ distance\ from\ present\ coordinates} \quad (1)$$

The weight is determined by dividing the total number of coordinates by the number of current coordinates [28].

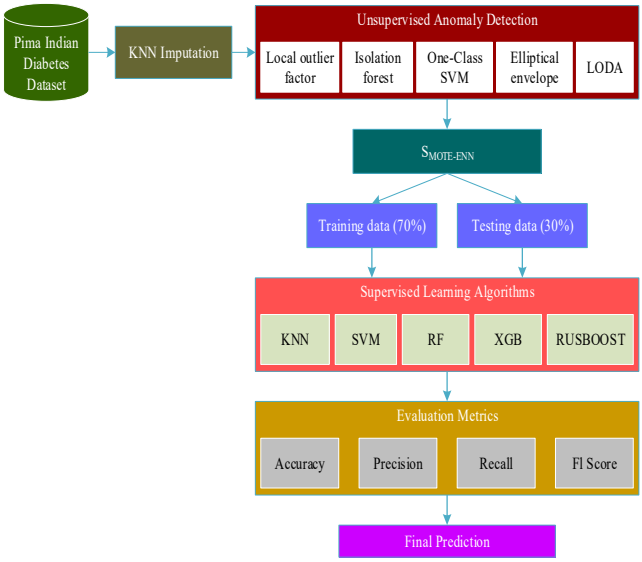

Figure 2. Architecture of proposed work.

The Algorithm 1 for the proposed approach is given below.

Algorithm 1: caption information

Input: Pima Indian Diabetes Dataset.

Output: Prediction of diabetes.

Step 1: Missing value imputation

All the zero values present in the dataset was replaced by NaN. Then KNN imputation was performed to impute the NaNs

Step 2: Unsupervised Anomaly Detection

LOF, iForest, OCSVM, eEnvelop and LODA was performed over the KNN imputed dataset and the anomalies were removed.

Step 3: Hybrid Sampling

After removal of anomalies from the dataset, Pima Indian Diabetes Dataset already being unbalanced becomes more imbalanced hence hybrid sampling is done. SMOTE-ENN was used to perform hybrid sampling.

Step 4: Diabetes Prediction

Train the classifiers namely KNN, SVM, RF, XGB and RUSBoost on training data to make diabetes predictions.

Step 5: Evaluation

Assess the model's performance by applying test data to the suggested model. Accuracy, precision, recall and F1 score were the measures

used to critique the approach.

3.2. Unsupervised Anomaly Detection Algorithms

Five distinct unsupervised methods for identifying anomalies are employed in this work because the anomalous data are unknown and unlabeled. Local Outlier Factor (LOF), Isolation Forest (iForest), One-Class SVM (OCSVM), Elliptic Envelope (eEnvelope), and Lightweight On-line Detector of Anomalies (LODA) are the five unsupervised algorithms.

3.2.1. Local Outlier Factor

Breunig et al. [29] proposed the local outlier factor algorithm. The algorithm is driven by k-NN (k-nearest neighbors) because, as its name implies, it evaluates each data point's degree of isolation from its neighbors. The intensity to which a point qualifies as an outlier can be determined using LOF. To calculate this local density approximation, K-nearest neighbors are utilized. Data points within the regions exhibiting lower density are classified as outliers or assigned a higher level of outlierness.

3.2.2. Isolation Forest (iForest)

In [30], Liu FT et al. suggested a technique known as Isolation Forest. A collection of iTrees is generated by iForest for a specific data set. Anomalies are those occurrences on the iTrees with short average path lengths. The size of the sample and the quantity of trees that need to be constructed are the only two factors considered by this method. With a small sub-sampling size, iForest can accomplish high detection performance with excellent effectiveness, and with fewer trees, its detection performance quickly converges.

3.2.3. One-Class SVM (OCSVM)

A specific type of Support Vector Machine (SVM) that is trained solely on data collected from a single class is known as a One-Class Support Vector Machine (One-Class SVM) [31]. Without needing a representative sample of outliers, it is frequently used for abnormal values detection or novelty detection, where the objective is to find records that differ substantially from the general trend of the data. The One-Class SVM algorithm aims to determine which hyperplane best divides the normal data points from the origin, thereby effectively enclosing the normal data within a narrow boundary. The decision boundary is the hyperplane that is acquired during the training process. Any data points outside this range are called anomalies or outliers. Like conventional SVMs, One-Class SVMs can project the data into higher-dimensional space to find a more effective separating hyperplane. This is accomplished by applying the kernel trick.

3.2.4. Elliptical Envelope (eEnvelope)

Anomaly or outlier detection is accomplished with the Elliptical Envelope model. If the data has a Gaussian distribution, this model performs better. It aims to draw an ellipse that comprises the majority of the data instances. Data points that fall noticeably outside of the ellipse are referred to as anomalous or outliers. The elliptic envelope method estimates the size and shape of the ellipse using the FAST minimum covariance determinant (FAST-MCD) [32].

3.2.5. Lightweight On-Line Detector of Anomalies (LODA)

Tomáš Pevný [33] proposed an outlier detection technique called the lightweight online detector of anomalies (LODA). LODA is an ensemble of weak estimators that is both simple and sophisticated, resulting in a robust and quick anomaly detection model. The main benefits of this model are its robustness to missing values, speed, simplicity, and capacity to explain anomaly causes. It also offers the option of online training. One-dimensional histograms built on sparse random projections make up LODA. LODA can process large datasets with relatively little time complexity using simple one-dimensional histograms made possible by random light projects. In addition, samples with missing features can be evaluated and even used in the training process by cleverly utilizing their sparsity.

3.3. Hybrid Sampling

Hybrid sampling techniques integrate oversampling and under-sampling techniques to take into consideration both the disproportionately represented minority class and the dominant majority class. Overfitting can occasionally result from oversampling. For this reason, it is best to clean the data by under-sampling before oversampling. SMOTE is the oversampling method used in the majority of these

hybrid strategies [34]. In this work, the Synthetic Minority Oversampling Technique and Edited Nearest Neighbor (SMOTE-ENN) are used. In order to function, SMOTE creates artificial examples in the minority class. In order to do this, it creates new samples at a specific point along a line that it draws between a few examples that are close to one another in the feature space. The way that ENN, an under-sampling technique, operates is by eliminating examples whose class label is different from the majority class label of the examples that are closest to it. This enhances the dataset's quality and helps to remove noisy examples [12]. The Algorithm 2 for SMOTE-ENN is given below.

Algorithm 2. for SMOTE-ENN

Input: Pima Indian Diabetes Dataset.

Output: Balanced Dataset.

SMOTE

mn ← minority class

mc ← majority class

Step 1: Select data at random from the non-dominant groups.

Step 2: Nieghbors= KNN (Find the k closest neighbors to it)

n= Determine one of the neighbors at random

Synthetic_values = [(original_values -n) * random (0,1)] + original_values

Step 3: Include the artificial sample in the dataset.

ENN

Step 4: Determine the k closest neighbors of each sample in the dataset.

Step 5: Remove the sample from the dataset if the majority class samples are more prevalent among the closest neighbors.

Step 6: Continue doing this until the predetermined number of iterations is reached or no more samples are removed.

3.4. Supervised Machine Learning Algorithm for Diabetes Prediction

3.4.1. K-Nearest Neighbor

KNN is the fundamental classification algorithm. A non-continuous function can be approximated using the K number of nearest classifiers. It builds a plane with the available training points to categorize them and computes the separation between the target and trained points. It counts K neighbors (based on the dataset) and classifies them or finds the mean of the closest neighbors to predict the values using majority voting [35].

3.4.2. Support Vector Machine

The Support Vector Machine (SVM) performs the dual roles of classifier and regressor. This algorithm determines a decision boundary that can split a space of N dimensions into classes, making it easier to place new data points in the appropriate class. The hyperplane is the ideal decision boundary, and the data points are support vectors. SVMs find the hyperplane with the most significant margin between classes [34,35].

3.4.3. Random Forest

An ensemble learning technique called Random Forest (RF) uses averaging to improve prediction accuracy and reduce overfitting. To do this, different Decision Trees are trained on different dataset subgroups. To classify a test object, RF builds a set of decision trees using randomly chosen sections of the training set. The final class is then determined by adding the votes from each decision tree. The technique will broaden the variety of its classifiers due to the features' random distribution, enhancing the model's prediction performance [34,35].

3.4.4. Extreme Gradient Boosting

Through the iterative development of a set of weak learners, XGB produces a reliable and accurate prediction model. In essence, XGBoost constructs a sturdy predictive model by joining predictions from multiple weak learners, typically decision trees. It employs a boosting technique whereby each weak

learner corrects the errors of their predecessors to produce an incredibly accurate ensemble model [35].

3.4.5. RUSBoost

For RUSBoost to function, weak learners typically use decision trees that are iteratively trained on various dataset iterations in which the minority class is randomly under and oversampled. Using this procedure, the dataset's imbalance is lessened, and the model can concentrate more on accurately classifying instances from the minority class. RUSBoost uses a weighted voting scheme to fuse the predictions of several weak learners trained to determine the final classification. More reliable and accurate classifiers are usually produced by combining the predictions of weak learners, particularly when dealing with unbalanced datasets. RUSBoost is appropriate and valuable for classification issues, particularly when handling unbalanced data [36].

4. Dataset, Evaluation Metrics and Results

4.1. Description of Dataset

The publicly accessible Pima Indian Diabetes Dataset is used to assess the suggested strategy. Women of Pima Indian descent who were 21 years of age or older had a diabetes test. The outcome variable is one of nine variables in the dataset. Pregnancies, Blood Pressure, Insulin, Skin Thickness, BMI, Diabetes Pedigree Function, and Age are the eight independent variables. There were 500 cases (65.1%) in class '0' and 268 cases (34.9%) in class '1' [10]. The dataset unbalance and contains 268 anomalies [12].

4.2. Evaluation Metrics

A model's performance is critiqued by several widely used assessment criteria. These include accuracy, precision, recall, and the F1 score. These metrics are computed using the following components of a confusion matrix: true negative (TN), false negative (FN), false positive (FP), and true positive (TP).

True Positive (TP): This happens when the trained model accurately diagnoses a record as having diabetes and the record's actual label reflects the same.

False Positive (FP): This occurs when a record's actual label is nondiabetic, but the trained model mistakenly predicts it to be diabetic.

True Negative (TN) occurs when a record's actual label is nondiabetic, and the trained model correctly predicts it to be nondiabetic.

False Negative (FN): This happens when a record is labeled as diabetic, but the trained model incorrectly predicts it to be nondiabetic.

Every evaluation metric used in this work is given as Equations (2)–(5).

Accuracy: As indicated by the formula below, this broad metric gives a ratio of correctly predicted instances to all instances in the dataset.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Precision: It, also known as favorable predictive value (PPV), can be mathematically represented as follows. It is the ratio of correct predictions to the overall right values, including both true and false predictions.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Recall: The proportion of correctly predicted values to the total of correctly predicted positive values and incorrectly predicted negative values is the mathematical expression for recall.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

F1 Score: Recall and precision are combined into one metric called the F1 score.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

4.3. Results

In this section, the results of diabetes prediction under four conditions are discussed to examine the importance of data quality before training any ML models. Figure 3 shows the results of diabetes prediction without anomaly detection and data balancing. The KNN imputed Pima Indian Diabetes

dataset was trained using five supervised algorithms. From all the five ML models, KNN, RF, and RUSBoost showed an accuracy of 0.75. RF and KNN had precisions of 0.72 and 0.70. Recall and F1 scores of RUSBoost were 0.72 and 0.69. Overall, the performance of RUSBoost was better without anomaly detection and balancing since RUSBoost performs random under-sampling by default.

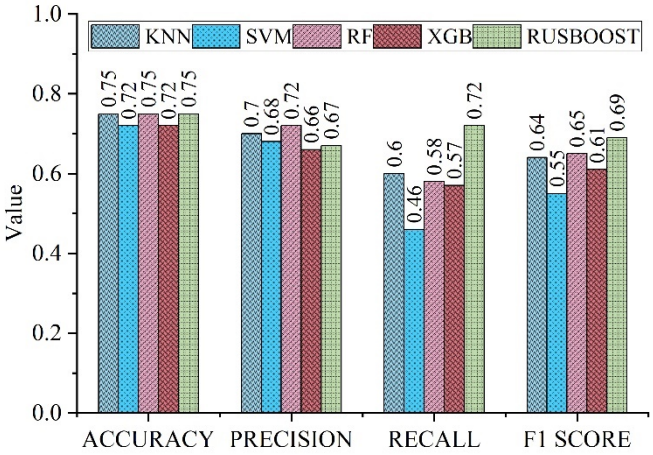

Figure 3. Performance measure of diabetes prediction with unbalanced dataset and without anomaly detection.

Figure 4 shows the results of diabetes prediction with hybrid sampling over KNN imputed dataset without anomaly detection. On performing SMOTE ENN, the overall performance of all the five ML models has been increased. XGB showed the accuracy of 0.86 and KNN reached the precision of 0.89, recall and F1 score of XGB was 0.88 and 0.87. Comparatively XGB's performed better with hybrid sampling.

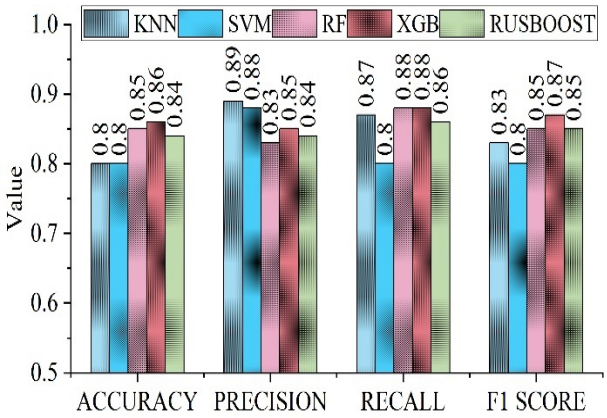

Figure 4. Performance measure of diabetes prediction with balanced dataset and without anomaly detection.

Table 1 depicts the outcomes with KNN imputation, anomaly detection and removal, no balancing of dataset and five supervised algorithms for diabetes prediction. The results are discussed in five cases.

In case 1: After performing LOF and removal of outliers, five ML models were applied to detect diabetes, out of which LOF + KNN produced 80% accuracy, 0.70 precision, recall of 0.64 and F1 Score of 0.67 followed by LOF + RF which shows the accuracy of 78%, precision of 0.69, recall of 0.66 and F1 score of 0.68.

In case 2: iForest anomaly detection and removal of anomalies were performed and it was found that iForest + SVM gave the accuracy and precision of 82% and 0.74, recall of 0.57 and F1 score of 0.65 followed by iForest + KNN gave the accuracy of 70% , 0.82 precision, recall of 0.65 and F1 score of 0.72.

In case 3: OCSVM+SVM produced the accuracy of 82%, precision of 0.85, recall of 0.63 and F1 score of 0.72 and OCSVM + RUSBoost attained an accuracy of 80%, precision of 0.65, recall of 0.83 and F1 score of 0.73.

In case 4: eEnvelope is the anomaly detection technique used to detect and remove anomalies. The accuracy and precision of eEnvelope + RUSBoost was 77% and 0.64, recall and F1 score was 0.80 and 0.71 and eEnvelope + SVM had the highest accuracy of 80%, precision of 0.79, recall of 0.60, F1 score of 0.68.

In case 5: LODA + RUSBoost performed with an accuracy of 75%, precision of 0.63, recall of 0.71, F1 score of 0.67 then LODA + SVM gave the accuracy of 79%. The performance of the classifiers dropped down when compared to condition 1 and 2.

Table 1. Results of diabetes prediction with anomaly detection in unbalanced dataset.

Techniques	Accuracy	Precision	Recall	F1 Score
LOF + KNN	80	0.70	0.64	0.67
LOF + SVM	75	0.76	0.50	0.60
LOF + XGB	72	0.60	0.59	0.59
LOF + RF	78	0.69	0.66	0.68
LOF + RUSBoost	75	0.61	0.75	0.67
iForest + KNN	70	0.82	0.65	0.72
iForest + SVM	82	0.74	0.57	0.65
iForest + XGB	71	0.37	0.36	0.36
iForest + RF	75	0.30	0.32	0.31
iForest + RUSBoost	76	0.57	0.78	0.66
OCSVM + KNN	78	0.72	0.45	0.55
OCSVM + SVM	82	0.85	0.63	0.72
OCSVM + XGB	78	0.68	0.64	0.66
OCSVM + RF	79	0.73	0.57	0.64
OCSVM + RUSBoost	80	0.65	0.83	0.73
eEnvelope + KNN	74	0.64	0.54	0.59
eEnvelope + SVM	80	0.79	0.60	0.68
eEnvelope + XGB	75	0.66	0.66	0.66
eEnvelope + RF	78	0.73	0.66	0.69
eEnvelope + RUSBoost	77	0.64	0.80	0.71
LODA + KNN	77	0.75	0.41	0.53
LODA + SVM	79	0.69	0.57	0.62
LODA + XGB	73	0.67	0.47	0.55
LODA + RF	76	0.75	0.49	0.60
LODA + RUSBoost	75	0.63	0.71	0.67

Table 2 shows the results of our proposed approach with KNN imputation, anomaly detection and removal of outliers, balancing of dataset and diabetes prediction. The results of this approach are discussed in five cases.

In case 1: In the LOF+SMOTE-ENN + ML Classifiers, 98.83% accuracy, 0.87 precision, 0.96 recall, and 0.91 F1 score were obtained by LOF + SMOTE-ENN + KNN, and 96% accuracy, 0.95 precision, 0.98 recall, and 0.96 F1 score were obtained by LOF + SMOTE-ENN + RF.

In case 2: The study evaluated iForest + SMOTE-ENN + ML classifiers. iForest + SMOTE-ENN + RF performed well, achieving 99% accuracy, 0.99 precision, 0.99 recall, and 0.99 F1 score. iForest + SMOTE-ENN + XGB followed closely behind, achieving 97% accuracy, 0.97 precision, 0.97 recall, and 0.97 F1 score.

In case 3: OCSVM + SMOTE-ENN +ML Classifiers were carried out. OCSVM + SMOTE-ENN+RF and OCSVM+SMOTE-ENN + RUSBoost attained the accuracy and precision of 96.80%, 0.94 and 96% and 0.94, recall and F1score of 0.99 and 0.97.

In case 4: eEnvelope + SMOTE-ENN + ML classifier was performed where eEnvelope + SMOTE-ENN + KNN gave the highest accuracy of 98%, precision and recall of 0.91 and 0.94 and 0.92 F1 score. eEnvelope + SMOTE-ENN + RF and eEnvelope + SMOTE-ENN + XGB obtained 97% accuracy, 0.96 precision,0.99 recall and 0.97 F1 score.

In case 5: The LODA + SMOTE-ENN + KNN classifier demonstrated the highest accuracy of 97.57%, 0.86 precision, 0.92 recall, and 0.89 F1 score among the LODA + SMOTE-ENN + ML classifiers that were tested. 92% accuracy, 0.92 precision, 0.92 recall, and 0.92 F1 score were obtained by LODA + SMOTE-ENN + RUSBoost.

Figure 5 depicts the maximum accuracy attained in four conditions. From the figure, it's clear that there was a progression in the performance when the dataset's quality was increased. In condition one, the accuracy reached 75%, then in condition 2, the accuracy increased to 86%; in condition 3, the

accuracy reduced to 82%, and in condition 4, the accuracy raised to 99%. Also, it is notable that the suggested approach produces better outcomes in comparison with the numerous current schemes, as shown in Table 3. The first best-performed approach is iForest + SMOTE-ENN + RF, followed by LOF + SMOTE-ENN + KNN and eEnvelope + SMOTE-ENN + KNN.

Table 2. Results of diabetes prediction with anomaly detection in balanced dataset.

Techniques	Accuracy	Precision	Recall	F1 Score
LOF + SMOTE-ENN + KNN	98.83	0.87	0.96	0.91
LOF + SMOTE-ENN + SVM	90.69	0.87	0.94	0.90
LOF + SMOTE-ENN + XGB	95	0.94	0.97	0.95
LOF + SMOTE-ENN + RF	96	0.95	0.98	0.96
LOF + SMOTE-ENN + RUSBoost	95	0.93	0.98	0.95
iForest + SMOTE-ENN + KNN	93	0.91	0.94	0.92
iForest + SMOTE-ENN + SVM	91	0.89	0.95	0.92
iForest + SMOTE-ENN + XGB	97	0.97	0.97	0.97
iForest + SMOTE-ENN + RF	99.23	0.99	0.99	0.99
iForest + SMOTE-ENN + RUSBoost	95	0.95	0.95	0.95
OCSVM + SMOTE-ENN + KNN	96.46	0.84	0.99	0.91
OCSVM + SMOTE-ENN + SVM	91	0.90	0.95	0.92
OCSVM + SMOTE-ENN + XGB	96	0.94	0.98	0.96
OCSVM + SMOTE-ENN + RF	96.80	0.94	0.99	0.97
OCSVM + SMOTE-ENN + RUSBoost	96	0.94	0.99	0.97
eEnvelope + SMOTE-ENN + KNN	98	0.91	0.94	0.92
eEnvelope + SMOTE-ENN + SVM	91	0.91	0.95	0.92
eEnvelope + SMOTE-ENN + XGB	97	0.96	0.99	0.97
eEnvelope + SMOTE-ENN + RF	97.08	0.96	0.99	0.97
eEnvelope + SMOTE-ENN + RUSBoost	95	0.94	0.97	0.95
LODA + SMOTE-ENN + KNN	97.57	0.86	0.92	0.89
LODA + SMOTE-ENN + SVM	88	0.89	0.90	0.90
LODA + SMOTE-ENN +XGB	91	0.90	0.93	0.92
LODA + SMOTE-ENN + RF	91.24	0.89	0.94	0.92
LODA + SMOTE-ENN + RUSBoost	92	0.92	0.92	0.92

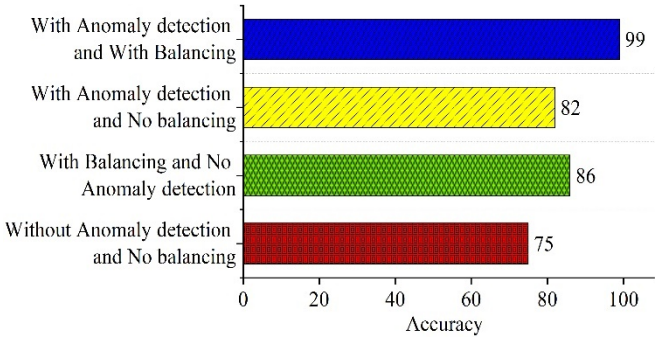

Figure 5. Overall comparison of highest accuracies reached in four conditions.

Table 3. Comparison of results of diabetes prediction with other methods.

Techniques	Accuracy	Precision	Recall	F1 Score
KNN [15]	76%	0.78	0.76	0.76
SVM [15]	75	0.78	0.75	0.76
Random Forest [15]	78	0.78	0.78	0.78
XGBoost [15]	78	0.78	0.78	0.78
PMSGD + (PIDD) [17]	79.55	0.79	0.56	0.65
PMSGD + (PIDD + GA) [17]	79.55	0.791	0.563	0.65
PMSGD + (PIDD + SM) [17]	82.12	0.80	0.85	0.83
PMSGD + (PIDD + SM + GA) [17]	80.19	0.77	0.85	0.81
eEnvelope + SMOTE-ENN + KNN (Proposed approach)	98	0.91	0.94	0.92
LOF + SMOTE-ENN + KNN (Proposed approach)	98.83	0.87	0.96	0.91
iForest + SMOTE-ENN + RF (Proposed approach)	99.23	0.99	0.99	0.99

The proposed approach handles data quality issues like zero values, anomaly detection, and unbalanced datasets. The proposed approach was assessed under four conditions. In the first condition, the diabetes prediction results showed that the maximum accuracy and F1 Score attained was 75% and 0.69 without anomaly detection and balancing dataset. In the second condition, the highest accuracy and F1 Score reached 86% and 0.87 with a balanced dataset and without anomaly detection. In the third condition, the accuracy and F1 Score were 82% and 0.73. The outcomes dropped because, after anomaly detection, the outliers were removed, and the dataset became more unbalanced. In the fourth condition, the highest accuracy and F1 Score reached 99.23% and 0.99 with anomaly detection and balancing. Hence, the above results demonstrate the importance of anomaly detection and balancing datasets in making predictions using ML models.

5. Conclusions

Diabetes may be a factor in the decline in both the length and quality of life. Predicting this lifelong condition earlier may lower the risk and adverse effects associated with multiple diseases. In this paper, a new approach for diabetes prediction addressing the issues associated with dataset quality has been assessed under four conditions, which include diabetes prediction using ML models with no balancing and no anomaly detection, a balanced dataset and without anomaly detection, then with anomaly detection and no balancing and last is with anomaly detection and balancing dataset. Assessment criteria's such as accuracy, precision, recall, and f1 score are used to evaluate the four conditions, and the results showed improvement after the dataset's issues with quality were fixed. iForest + SMOTE-ENN + RF gained the highest accuracy, precision, recall and f1 score of 99.23%, 0.99,0.99 and 0.99. The performance was dropped in third condition due to the removal of outliers, hence in future the outliers could be replaced with appropriate values using best working imputation methods. Furthermore, the proposed approach can be tested against various datasets to diagnose other diseases.

Author Contributions
All the authors contributed equally. All authors have read and agreed to the published version of the manuscript.

Funding
This research received no external funding.

Conflict of Interest Statement
Author declares no conflict of interest

Data Availability Statement
Dataset is available online: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

References

1. Pareek, Naveen Kumar, Deepika Soni, and Sheshang Degadwala. "Early Stage Chronic Kidney Disease Prediction using Convolution Neural Network." 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC). IEEE, 2023.
2. Roglic, Gojka. "WHO Global report on diabetes: A summary." International Journal of Noncommunicable Diseases 1.1 (2016): 3-8.
3. World Health Organization. "WHO Global report on diabetes." (2016).
4. Zhuhadar, Lily Popova, and Miltiadis D. Lytras. "The Application of AutoML Techniques in Diabetes Diagnosis: Current Approaches, Performance, and Future Directions." Sustainability 15.18 (2023): 13484.
5. Yadav, Pooja, et al. "Exploring Hyper-Parameters and Feature Selection for Predicting Non-Communicable Chronic Disease Using Stacking Classifier." IEEE Access (2023).
6. Gong, Youdi, et al. "A survey on dataset quality in machine learning." Information and Software Technology (2023): 107268.
7. Chatterjee, Ayan, and Bestoun S. Ahmed. "IoT anomaly detection methods and applications: A survey." Internet of Things 19 (2022): 100568.
8. Wu G, Chang EY. KBA: kernel boundary alignment considering imbalanced data distribution. IEEE Trans Knowl Data Eng.2005; 6:786-795.
9. Fernández, Alberto, et al. Learning from imbalanced data sets. Vol. 10. Cham: Springer, 2018.
10. Šabić, Edin, et al. "Healthcare and anomaly detection: using machine learning to predict anomalies in heart rate data." AI & SOCIETY 36.1 (2021): 149-158.
11. Torkey, Hanaa, et al. "Diabetes classification application with efficient missing and outliers' data handling algorithms." Complex & Intelligent Systems (2022): 1-17. 20-21
12. Samariya, Durgesh, et al. "Detection and explanation of anomalies in healthcare data." Health Information Science and Systems 11.1 (2023): 20.

13. Muntasir Nishat, Mirza, et al. “A comprehensive investigation of the performances of different machine learning classifiers with SMOTE-ENN oversampling technique and hyperparameter optimization for imbalanced heart failure dataset.” Scientific Programming 2022 (2022): 1-17
14. Oladimeji, Oladosu Oyebisi, and Olayanju Oladimeji. “Predicting survival of heart failure patients using classification algorithms.” JITCE (Journal of Information Technology and Computer Engineering) 4.02 (2020): 90-94.
15. Tasin, Isfafuzzaman, et al. “Diabetes prediction using machine learning and explainable AI techniques.” Healthcare Technology Letters 10.1-2 (2023): 1-10.
16. Wang, Qian, et al. “DMP_MI: an effective diabetes mellitus classification algorithm on imbalanced data with missing values.” IEEE access 7 (2019): 102232-102238.
17. Azad, Chandrashekhar, et al. “Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus.” Multimedia Systems (2021): 1-19.
18. Ramesh, J., Aburukba, R., Sagahyroon, A.: A remote healthcare monitoringFramework for diabetes prediction using machine learning. HealthcareTechnol. Lett. 8, 45–57 (2021)
19. Nnamoko, Nonso, and Ioannis Korkontzelos. “Efficient treatment of outliers and class imbalance for diabetes prediction.” Artificial intelligence in medicine 104 (2020): 101815.
20. Fitriyani, Norma Latif, et al. “Development of disease prediction model based on ensemble learning approach for diabetes and hypertension.” Ieee Access 7 (2019): 144777-144789.
21. Srivastava, Anand Kumar, Yugal Kumar, and Pradeep Kumar Singh. “Hybriddiabetes disease prediction framework based on data imputation and outlier detection techniques.” Expert Systems 39.3 (2022): e12785.
22. Godase, Uma R., and Darshan V. Medhane. “OptDCE: An optimal and diverse classifier ensemble for imbalanced datasets.” International Journal of Computer Information Systems and Industrial Management Applications 14 (2022): 11-11.
23. Sharma, Vibhuti, Anu Bajaj, and Ajith Abraham. “Machine Learning Techniques for Electronic Health Records: Review of a Decade of Research.” International Journal of Computer Information Systems & Industrial Management Applications 15 (2023).
24. Rathod, S. R., and C. Y. Patil. “Linear and non-linear HRV features for the prediction of heart disease among smokers: a predictive evaluation of machine learning model.” International Journal of Computer Information Systems and Industrial Management Applications 12 (2020): 9-9.
25. Wang, Shujuan, et al. “Research on expansion and classification of imbalanced data based on SMOTE algorithm.” Scientific reports 11.1 (2021): 24039.
26. Alnowaiser, Khaled. “Improving Healthcare Prediction of Diabetic Patients Using KNN Imputed Features and Tri-Ensemble Model.” IEEE Access (2024).
27. Aruna Devi B, and Karthik N. “Data Imputation using Correlation-based Machine Learning Algorithms.” Intelligent Systems Design and Applications: 23rd International Conference on Intelligent Systems Design and Applications (ISDA 2023), Volume 2. Springer International Publishing, 2024.
28. Breunig MM, Kriegel H-P, Ng RT, Sander J. LOF: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD international conference on management of data. 2000, p. 93–104. (An ensemble)
29. Liu FT, Ting KM, Zhou Z-H. Isolation Forest. In: 2008 Eighth IEEE international conference on data mining. IEEE; 2008, p. 413–22.
30. Schölkopf B, Williamson RC, Smola A, Shawe-Taylor J, Platt J. Support vector method for novelty detection. In: Advances in neural information processing systems. Vol. 12. 1999.
31. Ashrafuzzaman, Mohammad, et al. “Elliptic envelope-based detection of stealthy false data injection attacks in smart grid control systems.” 2020 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2020.
32. Pevný, Tomáš. “Loda: Lightweight on-line detector of anomalies.” Machine Learning 102 (2016): 275-304.
33. Susan, S., and A. Kumar. “The balancing trick: Optimized sampling of imbalanced datasets—A brief survey of the recent State of the Art. Engineering Reports, 3 (4), e12298.” (2021).
34. Tasin, Isfafuzzaman, et al. “Diabetes prediction using machine learning and explainable AI techniques.” Healthcare Technology Letters 10.1-2 (2023): 1-10.
35. Dutta, Aishwariya, et al. “Early prediction of diabetes using an ensemble of machine learning models.” International Journal of Environmental Research and Public Health 19.19 (2022): 12378.
36. Seiffert, Chris, et al. “RUSBoost: A hybrid approach to alleviating class imbalance.” IEEE transactions on systems, man, and cybernetics-part A: systems and humans 40.1 (2009): 185-197.