2023 IEEE Symposium on Computers and Communications (ISCC): 28th IEEE Symposium on Computers and Communications - 1st Workshop on Machine Learning from Class Imbalanced Data

The Effect of Feature Selection on Diabetes Prediction using Machine Learning

Rania Alhalaseh
Information Technology Faculty
Mutah University
Karak, Jordan
halaseh@mutah.edu.jo

Dhuha Ali Ghani AL-Mashhadany
Information Technology Faculty
Mutah University
Karak, Jordan
dhuhaghani90@gmail.com

Mohammad Abbadi
Information Technology Faculty
Mutah University
Karak, Jordan
abbadi@mutah.edu.jo

***Abstract*—Diabetes arises from inadequate insulin production or ineffective response to insulin in the body, leading to long-term health complications such as cardiovascular disorders and organ dysfunction. This study aims to enhance the accuracy of diabetes prediction using machine learning techniques. Two datasets, namely the Pima Indians Diabetes and Mendeley datasets, were utilized to evaluate the performance of various machine learning classifiers. The Mendeley dataset presented the challenge of class imbalance, which was effectively addressed by employing a novel data balancing technique called Random Data Partitioning with Voting Rule (RDPVR), resulting in improved accuracy in diabetic prediction. Feature selection methods, including Recursive Feature Elimination (RFE), Analysis of Variance (ANOVA), and Step Forward (SF), were applied to eliminate irrelevant information and optimize the efficiency of the machine learning algorithms. Logistic Regression (LR), Nave Bayes (NB), and Random Forest (RF) classifiers, along with ensemble soft voting, were utilized. The experimental results indicated that Recursive Feature Elimination combined with ensemble soft voting achieved the highest accuracy of 97% using the Mendeley dataset and 81% using the Pima Indians Diabetes Dataset.**

***Index Terms*—diabetes, imbalance data, RDPVR, machine learning, feature selection.**

I. Introduction

Diabetes is considered a chronic metabolic disorder affecting a large number of human beings globally. Diabetes disease takes place when the body does not produce a convenient quantity of insulin or does not respond well to insulin. If diabetes is not properly treated long-term health troubles appear which comprise: heart issues, skin and liver complications, nerve harm, etc. [1]. There are different kinds of diabetes:

- Type 1 diabetes: when the body is unable to produce insulin. In this case, patients depend on insulin injections every day to control their blood sugar levels.
- Type 2 diabetes: impacts the way body utilizes insulin, where patients' bodies still produce insulin compared to Type 1.
- Type 3 diabetes (Gestational): appears during pregnancy because the body cannot produce enough insulin. This type of diabetes does not happen in all females and stabilizes after childbirth [2].

Type 2 diabetes accounts for 87% - 91% of all diabetes types. Therefore, the prediction of Type 2 is key to controlling the global spread of diabetes [3], [4]. In recent years, machine learning algorithms have been widely used in the medical domain.

For diabetes prediction, this work used two datasets: First, Pima Indian Diabetes Dataset (PIDD) is originally from the National Institute of Diabetes and Digestive and Kidney Diseases; it has 768 samples where 500 are non-diabetes and 268 are diabetes patient [5]. Various machine learning algorithms and ensemble approaches have been used on this dataset for the prediction of disease but none of them could reach a high accuracy [6].

The second dataset which is considered a new dataset from the laboratory of Medical City Hospital and the Specializes Center for Endocrinology and Diabetes (Al-Kindy Teaching Hospital in Iraq), is called Mendeley dataset [7]. This dataset is imbalanced which leads to wrong predictions therefore the accuracy of the machine learning algorithm will be incorrect. Such as the accuracy might be very high but in fact, this is not right [8].

In literature different techniques were used to solve the problem of imbalanced data, this work uses a new technique called Random Data Partition Using Voting Rules (RDPVR) [9].

Moreover, for better efficiency results of machine learning algorithms, the feature selection step is used to identify and remove as much of the irrelevant information as possible [10].

This work used different feature selection methods: Recursive Feature Elimination (RFE), Analysis Of Variance (ANOVA), Step Forward Feature Selection methods. Also, three machine learning algorithms were used: Logistic regression (LR), Nave Bayes (NB), Random Forest (RF) and ensemble soft voting of three classifiers. The objective of this work is summarized as follows:

1) studying the effect of feature selection on diabetes prediction and which feature is the most important that affects the prediction process.
2) using RDPVR method to balance Mendeley dataset in order to avoid the drawbacks of using imbalanced datasets and improve the accuracy of diabetes prediction.
3) evaluating and testing the proposed model on different datasets: PIDD dataset and Mendeley dataset where the new approach for data balancing was used.

II. Literature Review

For diabetes prediction, plenty of methods have been developed to obtain the best results. The following section shows different studies on diabetes prediction.

A. Type 2 diabetes prediction using machine learning

In a study done by [11] used different classification algorithms such as Naive Bayes, Neural Networks, SVM, and

2023 IEEE Symposium on Computers and Communications (ISCC) | 979-8-3503-0048-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISCC58397.2023.10218243

Decision Tree to predict diabetes. This study used PIDD dataset where the best results were obtained by SVM 76.8%.

[12] used ensemble soft voting of three classifiers which are RF, LR, NB to increase the performance of diabetes prediction. PIDD is used for checking the proposed work. The higher accuracy was recorded using the ensemble method 79.08%.

B. Using feature selection methods in type 2 diabetes prediction

[13] used two different datasets: PIDD and Vanderbilt dataset which are available in their website. This work used Univariate feature selection and ensemble method max voting and stacking. They used Nave Bayes, k-Nearest Neighbor, Decision Tree, SVM, and LR as based classifiers in the ensemble method. The higher accuracy achieved from the ensemble (Max Voting) was 78% using PIDD and 93% using the Vanderbilt dataset.

on the other hand, [14] used different classifiers like ANN, XGboost for diabetes prediction and RFE for the feature selection step. The dataset used in this work PIDD dataset XGboost has the highest accuracy 78.1%.

Along the same line of inquiry, [15] used Naive Bayes, DT, SVM classifiers to detect diabetes. PIDD was used in this research. The authors did not discuss the data preprocessing. The segmentation of the dataset is done by means of $10-folds$ cross-validation. The experiment results display that Naive Bayes produced the highest accuracy which reached 76.30%.

III. Proposed Model

The proposed model is shown in Figure 1 which includes: datasets, preprocessing, feature selection, and prediction process. The following subsections explain each step.

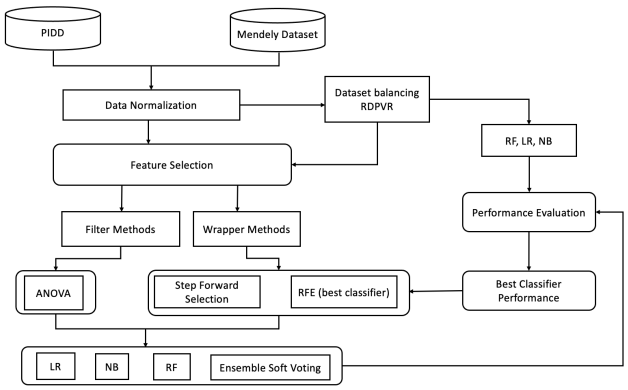


Fig. 1: The Proposed Model

A. Dataset

1) Pima Indian Diabetes Dataset (PIDD): PIDD is taken from the national institute of Diabetes and Digestive and Kidney Diseases (NIDDK). It is available from Kaggle and UCI data repositories [16]. The dataset was collected from females, all of them were randomly selected such that they are at least 21 years old of Pima Indian heritage. The dataset contains 768 instances and 9 attributes. The dataset has 268 diabetes patients and 500 non-diabetes patients. Table I describes all features in dataset [17].

TABLE I: PIDD Dataset Description [5]

No.	Feature	Description	Range
1	N- pregnancies	Pregnancy count of women	(017)
2	AGE	Age of female patient	$(21-81)$
3	BMI	Body mass function	$(0-17)$
4	Insuline	2-hour serum insulin	$(0-846)$
5	Skin thickness	Triceps skin fold thickness	$(0-99)$
6	PDF	Diabetes pedigree function	$(0.078-2.42)$
7	Glucose	Plasma glucose concentration	$(0-199)$
8	dbp	Diastolic blood pressure	$(0-122)$
9	Outcome	Diabetes diagnoses results	$(0,1)$

2) Mendeley Diabetic Dataset: The dataset is collected from the Iraqi society and obtained from the laboratory of Medical City Hospital and the Specializes Center for Endocrinology and Diabetes-Al-Kindy Teaching Hospital. Patients' files were taken and entered into the database to construct the diabetes dataset. The dataset contains 947 instances and 10 attributes. The dataset contains 103 instances non-diabetic, and 844 are diabetes. The data attributes are described in TableII.

TABLE II: Mendeley Dataset Description [7]

No.	Feature	Description	Range
1	Gender	Male or Female	
2	Age	Age of patients	$(20-69)$
3	Urea	Urea level	$(0.5-38.9)$
4	Creatinine	Creatinine Serum	$(48-80)$
5	BMI	$(19-47)$	
6	LDL	Low-Density Lipoprotein	$(0.3-9.9)$
7	HDL	High-Density Lipoprotein	$(0.2-9.9)$
8	VLDL	Very-Low-Density Lipoprotein	$(0.1-35)$
9	Triglycerides	Triglycerides are type of fat (lipid) blood	$(0.3-13.8)$
10	HbA1C	glucose level attached to hemoglobin	$(0.9-16)$
11	Class	Diabetic, no diabetic, pre-diabetic	

B. Data Preprocessing

Data Preprocessing is a crucial step in transforming data into a usable and proper format. The proposed model employed two approaches: data normalization and handling class imbalance.

1) Data Normalization: This method is used to conduct linear data transformation which involves a linear transformation of the original data with all attribute values ranging between $[0,1]$ [14].

2) Handling Class Imbalance: The problem of class imbalance happens when the dataset has very high proportions of one class compared to the other class(es). The first class is generally mentioned as the majority class, while the other is indicated to as the minority. The essential problem with class imbalance is the poor predictive performance for the minority class when unequal training sets are trained by classifiers [18], [19]. The minority class is considerable of importance because it represents positive cases that are rare in nature or costly to obtain. It is challenging especially when dealing with medical cases and disease prediction [20], [21].

Two major approaches for resolving an unbalanced dataset problem are: oversampling minority class instances and undersampling majority class ones [22], [23].

Oversampling is the most used approach where new samples are established from nothingness depending only on their sameness to one or more of the minoritys samples [24]. In theory, overfit synthetic datasets generate good machine-learning outcomes, but in fact, this is not always the case. Oversampling refers to the possibility that the created instances could exist in the actual world and belong to a majority class, regardless of how similar they are to the minority's instances. There are no tests for the validity of the synthesized instances or if they are adequate for training a model for real-world use, hence the sole methodology for validating an oversampling method's usefulness is its

classification accuracy metrics after the classification of the oversampled datasets [25].

On the other hand, the number of samples from the majority classes is reduced with under-sampling strategies [26]. Both methods have their hold set of problems [24]. This also gives positive results on paper, but in practice, the opposite is true [9].

This work uses Random Data Partitioning with Voting Rule (RDPVR) to avoid the deficiencies of both oversampling and undersampling approaches.

Random Data Partitioning with Voting Rule (RDPVR) based on generating a balanced learning process for imbalanced datasets where the imbalanced dataset is partitioned to smaller balanced sub-datasets. This is accomplished by selecting numerous majority samples from the same number of minority samples as random, disregarding all other majority samples, and keeping all minority samples for each sub-dataset. This method ensures that each sub dataset is balanced. After that, utilize every one of these sub-datasets individually, yielding various trained models that were all used to classify a simple voting conduct [9]. Figure 2 shows the RDPVR method. This work employed the same classifiers that were used to predict diabetes for the RDPVR technique.

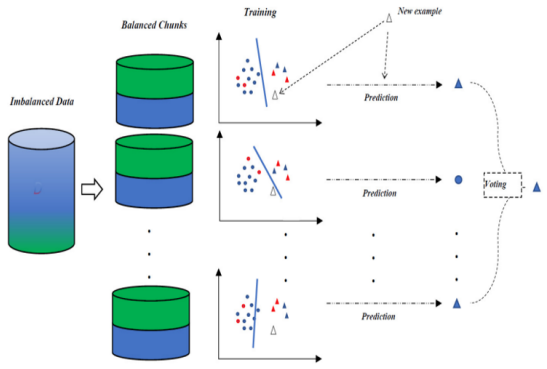


Fig. 2: Procedure of RDPVR Technique [9].

C. Feature Selection

The dataset could include numerous features, but not all of them contribute to the predictive process. The task at hand is to identify the most valuable features, which necessitates employing a dimensionality reduction technique. This approach reduces the dataset's size by eliminating less significant features. Referred to as the feature selection method, it is utilized to decrease the overall data size [27]. The aim is to enhance the performance of the learning algorithm and minimize the complexity of the data selection process.

Filter, wrapper, and embedded methods are the three types of traditional feature selection methods [28]. This work used filter and wrapper ones.

Filter Methods is considered one of the most traditional methods of feature selection [29], [30]. The statistical connection between a group of variables and the target variable is used to identify and rank the essential ones [31]. The filter method uses a threshold, a categorical variable, and the evaluation value of a single feature to determine the selection and examination of the features. A feature with a low evaluation value may be eliminated based on the evaluation result, but in some cases, a feature with a low evaluation value may have a well-combined classification effect, so the feature should not be chosen solely based on the single evaluation value, but rather the sorted classification effect [32]. Based on this approach Analysis of Variance (ANOVA) is regarded as a parametric technique that is used to link class mean for a certain dataset feature depending on computing the mean between different groups within a dataset [33], [34].

Wrapper Methods explore different feature subsets by leveraging a learning algorithm to guide the search process. This approach involves calculating the estimated accuracy of the learning algorithm for each potential feature and assessing its impact on the feature subset. Wrapper methods utilize the output of the classifier to identify the most suitable feature subset. One advantage of this approach is that it considers both the feature subset search and model selection simultaneously. However, a potential drawback is an increased risk of overfitting. In this study, the Step Forward and Recursive Feature Elimination (RFE) techniques are employed for feature selection [35], [36], [37].

1) Recursive Feature Elimination has an iterative procedure that constructs the model on all features and ranked the feature according to its important [38]. The feature with the shortest ranking criterion is deleted in the second phase since it has the smallest impact on performance [39]. After that, updated features are used to re-train the model and compute the feature importance again. This operation is repeated until the feature set is empty [40].
2) On the other hand, step-forward feature selection is a recursive selection method. It begins with the best fulfilling feature with the target. In the next step, it chooses the best-performing feature and combines it with the previously best-selected one. This is repeated till the addition of new features does not enhance the performance of the model. After each iteration, just the k best feature subsets are preserved to reduce the number of evaluations [41].

D. Classification Algorithms

The proposed model uses three algorithms: Logistic Regression, Naive Bayes, Random Forest classifier, and Ensemble soft voting.

Logistic Regression (LR) finds application in various fields, including the biological sciences. It is a statistical modeling technique that associates a set of explanatory factors with the probability of belonging to a particular category. The target variable in LR is binary, meaning it comprises data classified as either 0 or 1 [42]. The LR classifier offers several advantages, such as ease of use, efficient computation, and reduced risk of overfitting. However, it may struggle to capture intricate correlations between input and output variables [43].

Random Forest (RF) is considered one of the most famous machine learning classifiers in the medical domain with powerful performance. RF is an ensemble learning algorithm that depends on the concept of bagging or bootstrap aggregation. A bagging algorithm proves to be very effective to reduce the variance of a model by creating several subsets by randomly selecting samples from the training data. Each subset is then used to train decision trees in detached after that compute the averaged of all of these predictions and ensembled to make

a strong prediction [44]. This method gives good prediction results because it produces a forest of decision trees.

Nave Bayes (NB) classifier depends on Bayes theorem which assumes that no existing interconnection between the presence of one feature in a given class with another feature in the same class [45], [46]. It can be utilized to calculate the probability of a proposed diagnosis is right. NB is beneficial, for high dimensional data where each feature is estimated probability independently [47].

Voting Ensemble Approach refers to the method that combines many models to make a decision, which is particularly useful in supervised machine learning applications. The primary objective of ensemble learning is combining numerous models, the ensemble's overall prediction performance might be better than a single model because a single model's faults are likely to be compensated by another model [48]. Hard voting and soft voting are the most popular used schemes. Hard voting aims to appreciate the final class label by computing the majority of labels predicted by all classifiers [49]. Whereas soft voting aims to estimate the final class label, it computes the weighted total of the prediction probabilities of all classifiers for each class [50].

IV. Results

The first set of experiments includes the performance of ML classifiers using all features (without feature selection) on the used datasets. The second set includes using different feature selection methods to select important features. Also, a comparison between the proposed model and other state-of-the-art approaches is addressed.

Machine learning algorithms: Logistic regression, random forest, and Naive-Bayes, as well as an ensemble are utilized to evaluate the PIDD and Mendeley datasets for diabetes prediction. The training and testing processes are divided into $n-folds$, with n equal to 10. The data is divided into n equal folds in $n-folds$, and the trials run in $n-rounds$. Each cycle has $n-1folds$ for training and $1-fold$ for testing. As a result, each of the folds is utilized as a testing set in each cycle, allowing all of the data to be tested.

A. Experiments Using All Features

In the first set of experiments, the results of several classifiers using all features in the dataset (without using feature selection methods) will be presented.

As represented in Table III RF classifier gave a higher performance than other classifiers. In Mendeley dataset it records 95%, 91%,93%,91%, on the other hand, in PIDD record 80%, 78%,79%,78%, for precision, recall, F-score and accuracy respectively. The RF performed very well using all features in PIDD compared to other classifiers. This may happen because RF reduces overfitting in decision trees and helps to improve accuracy.

B. Experiments using Feature selection methods

The aim of feature selection is to determine important features in the dataset and neglect the unimportant ones in order to increase accuracy. This work focuses on using RFE, step forward for the feature selection step. Then using the following classifiers: LR, RF, NB, and Ensemble soft voting.

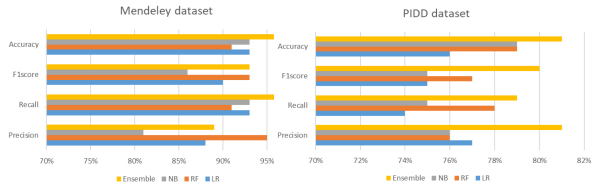
(a) Mendeley Dataset

Classifier	Precision	Recall	F1 Score	Accuracy
LR	89%	88%	87%	88%
RF	95%	91%	93%	91%
NB	79%	91%	83%	90%

(b) PIDD Dataset

Classifier	Precision	Recall	F1 Score	Accuracy
LR	79%	77%	77%	76%
RF	80%	78%	78%	78%
NB	76%	74%	74%	74%

TABLE III: Performance of Different Classifiers Before Feature Selection



(a) Mendeley Dataset (b) PIDD Dataset

Fig. 3: Performance of Classifiers after Using RFE

1) Recursive feature elimination (RFE): Figure 3a indicates the performance of the various classifiers using RFE feature selection method. There is a clear improvement in the performance when using Mendeley dataset with all features, and a simple improvement in performance when using PIDD. LR and NB recorded 93% as accuracy on Mendeley dataset and 91% when using RF. Ensemble soft voting recorded a higher accuracy of 97% which is better than that of a single model because in an ensemble model, the errors of a single model will likely be compensated by another model. Figure 3 shows the performance of classifiers after using REF on both datasets; Mendeley in Figure 3a and PIDD in Figure 3b.

On the other hand, table IV shows the cross-validation values of PIDD features using RFE feature selection method. The five best features of PIDD are Glucose, Diastolic, Age, Dia-pred, and BMI.

(a) Mendeley Dataset using RFE

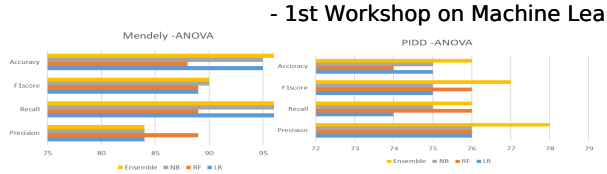
Feature	Rank	Selected
Gender	False	2
Age	True	1
Urea	False	3
Cr	True	1
HbA1c	True	1
Chol	True	1
TG	True	1
HDL	False	4
LDL	False	5
VLDL	False	6
BMI	True	1

(b) PIDD Dataset using RFE

Feature	Rank	Selected
N-preg	False	2
Glucose	True	1
Diastolic	True	1
Skin thickness	False	3
Insulin	False	4
BMI	True	1
Dia-pred	True	1
Age	True	1

TABLE IV: Ranking Features of Datasets using RFE

2) Analysis of Variance (ANOVA): ANOVA is a technique used for feature selection, which involves calculating the means between different groups within a dataset [33]. Figure 4 visually presents the results obtained after applying ANOVA to both datasets. Figure 4a specifically displays the scores of each feature in the Mendeley dataset after utilizing ANOVA for feature selection. Notably, BMI emerges as the feature with the highest f-score, indicating its significant weightage during the feature selection process. The top five features in the Mendeley dataset are BMI, HbA1c, Age, HDL,



(a) Mendeley Dataset (b) PIDD Dataset

Fig. 4: Performance of Classifiers after Using ANOV

and Chol [33].

As shown in Figure 4a the top three feature values in Mendeley dataset are BMI, HbA1c, Age. This is compatible with the medical domain. According to [51] the HbA1c value is used in the diagnosis of diabetes. The risk of having diabetes increases with the increase in BMI [52], [53] and with increases in age [54].

On the other hand, the top three features values in PIDD as shown in Figure 4b are: Glucose, Age, BMI. Also, this is compatible with medical domain because diabetes occurs because of the excess sugar level or glucose in our blood.

Moreover, the higher recorded accuracy is 96% by using the ensemble method and followed by 95% using NB classifier. Also, the Ensemble method and NB classifier recorded higher recall by 98%. Also, results of ANOVA feature selection method with PIDD are as follows: ensemble method recorded higher accuracy, Precision, Recall ,F1-score which are 78%, 76%,77%,76%. This is shown graphically in Figure 4b.

3) Step Forward Feature Selection: Based on this feature selection approach, the selected features are Age, Cr, HbA1c, LDL by using Mendeley dataset. It starts with HbA1c with a 97% score. The score increases when adding Cr feature and continually increases until reaches to 100% using HbA1c, Cr, AGE, and LDL.

The highest score using Mendeley dataset is 100% using Age, Cr, HbA1c, LDL features. This means that higher performance is achieved when using these features. On the other hand, in PIDD the feature selected are Glucose, BMI, Age, N-preg, Dia-pred, Skin-thickness, and Insulin. The score starts with 67% using the Glucose feature only and continuously increases until reaches 80% when using Glucose, BMI, Age, N-preg, Dia-pred, Skin-thickness, and Insulin features.

As shown in table V when using Mendeley dataset the accuracy is not high compared with other feature selection methods. This happens because Step Forward is unable to remove features that become non-useful after the addition of other features. The higher accuracy, precision, recall, and F-score recorded are 92%,79%,93%, and 84% using LR classifier respectively.

TableVI depicts the most significant features in Mendeley dataset which are HbA1c, BMI, and AGEA. It depicts the most significant features in PIDD: Glucose-conc, BMI, and AGE. These features are considered the most significant ones for diabetes prediction.

According to American Diabetes Association (ADA) and World Health Organization (WHO) the HbA1c value is used to diagnose diabetes [52]. This interprets the strong relation between HbA1c and outcome. Moreover, it is mentioned by [52] that The number of people with diabetes increases with the increase in BMI.

(a) Mendeley Dataset

Classifier	Precision	Recall	F1 Score	Accuracy
LR	79%	93%	84%	92%
RF	83%	84%	82%	84%
NB	71%	85%	75%	85%
Ensemble	79%	91%	83%	90%

(b) Mendeley Dataset

Classifier	Precision	Recall	F1 Score	Accuracy
LR	79%	76%	78%	76%
RF	75%	74%	75%	74%
NB	76%	73%	74%	74%
Ensemble	79%	79%	k79%	77%

(c) PIDD Dataset

TABLE V: Results of Diabetes Prediction after using Step Forward

(a) Mendeley Dataset

Feature Name	Importance
HbA1c	0.54
BMI	0.54
Age	0.41
TG	0.17
Chol.	79%
Gender	79%
VLDL	79%
Cr	79%
HDL	79%

(b) PIDD Dataset

Feature Name	Importance
HbA1c	0.54%
BMI	0.54%
Age	0.41%
TG	0.17%
Chol.	79%
Gender	79%
VLDL	79%
Cr	79%
HDL	79%

TABLE VI: Features's importance order of Mendeley and PIDD dataset

C. Comparison with Previous Studies

The proposed model is compared with previous studies for diabetes prediction, as shown in table VII. The importance of the proposed model is the use of the feature selection process where the results of prediction are done with and without this step. Moreover, it tests different classifiers on the new dataset Mendeley as well as shows the effect of using RDPVR on the prediction results.

TABLE VII: Comparison with Other Previous Research

Reference	Classifier	Dataset	Accuracy
[55]	LR, NB, RF, Ensemble soft voting	PIDD	Ensemble 79%
[13]	LR, ensemble method	PIDD	ensemble Max voting 77%
[14]	ANN, XGboost	PIDD	XGboost 78.1%
[15]	SVM, DT, NB	PIDD	NB 76.30%
[56]	LR, KNN, SVM,NB,DT, RF	PIDD	RF 75%
Our work	LR, NB, RF, Ensembele soft voting	PIDD, Mendeley	Ensemble: PIDD 81%, Mendeley 97%

CONCLUSIONS

This study focuses on two key aspects: the significance of the feature selection step and addressing the challenge of class-imbalanced datasets in diabetes prediction. The proposed model is evaluated using two datasets: PIDD and Mendeley. Notably, the Mendeley dataset serves as an example of a class-imbalanced dataset. To address this issue, the RDPVR approach is employed, which effectively mitigates the problems of overfitting and underfitting commonly associated with oversampling and undersampling methods. By utilizing the RDPVR approach, this work aims to overcome the challenges posed by class imbalance in the datasets used for diabetes prediction.

In this study, feature selection methods were employed to identify the most informative features in the datasets. Both filter methods and wrapper methods were utilized for this purpose. Among the filter methods, ANOVA was employed, while the wrapper methods included RFE and Step Forward selection. The machine learning algorithms used in the analysis were LR, NB, and RF classifiers. The results showed

that ensemble soft voting yielded improved performance in terms of prediction accuracy, indicating the effectiveness of combining multiple classifiers.

Based on the experimental results using different datasets, the highest accuracy achieved was 81% for the PIDD dataset and 97% for the Mendeley dataset. These results were obtained by applying RFE as the feature selection method and utilizing ensemble soft voting with three classifiers. From the feature selection analysis, it was found that the most influential features impacting the prediction results varied between the PIDD and Mendeley datasets. In the PIDD dataset, the Glucose, BMI, and AGE features were identified as the most significant. Conversely, in the Mendeley dataset, the HbA1c, BMI, and AGE features were found to have the highest impact on the prediction results.

These findings have significant implications for the medical field, as identifying the key features that contribute to accurate prediction models can lead to improved diagnostic and treatment approaches in the future.

REFERENCES

[1] H. Thakkar, V. Shah, H. Yagnik, and M. Shah, "Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis," *Clinical eHealth*, vol. 4, pp. 12–23, 2021.

[2] S. Joshi and P. Shetty, "Performance analysis of different classification methods in data mining for diabetes dataset using WEKA tool," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 3, pp. 1168–1173, 01 2015.

[3] J. Hou, Y. Sang, Y. Liu, and L. Lu, "Feature selection and prediction model for type 2 diabetes in the chinese population with machine learning," *Proceedings of the 4th International Conference on Computer Science and Application Engineering*, 2020.

[4] R. Canlas, "Data mining in healthcare: Current applications and issues," *School of Information Systems & Management, Carnegie Mellon University, Australia*, 2009.

[5] K. Website, "Pima indians diabetes database." Accessed on 2023.

[6] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," *Journal of Intelligent Learning Systems and Applications*, vol. 09, pp. 1–16, 2017.

[7] A. Rashid, "Diabetes dataset, mendeley data," 2020. Accessed on 2023.

[8] A. J. Mohammed, M. M. Hassan, and D. H. Kadir, "Improving classification performance for a novel imbalanced medical dataset using SMOTE method," *International Journal*, vol. 9, no. 3, pp. 3161–3172, 2020.

[9] A. B. Hassanat, A. S. Tarawneh, S. S. Abed, G. A. Altarawneh, M. Alrashidi, and M. Alghamdi, "RDPVR: Random data partitioning with voting rule for machine learning from class-imbalanced datasets," *Electronics*, vol. 11, no. 2, 2022.

[10] Y. Huang, P. Mccullagh, N. D. Black, and R. Harper, "Feature selection and classification model construction on type 2 diabetic patients' data," *Artificial intelligence in medicine*, vol. 41 3, pp. 251–62, 2007.

[11] S. Sossi Alaoui, B. Aksasse, and Y. Farhaoui, "Data mining and machine learning approaches and technologies for diagnosing diabetes in women," in *Big Data and Networks Technologies 3*, pp. 59–72, Springer, 2020.

[12] M. Alghamdi, M. Al-Mallah, S. Keteyian, C. Brawner, J. Ehrman, and S. Sakr, "Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The henry ford exercise testing (FIT) project," *PLOS ONE*, vol. 12, 07 2017.

[13] P. Rajendra and S. Latifi, "Prediction of diabetes using logistic regression and ensemble techniques," *Computer Methods and Programs in Biomedicine Update*, vol. 1, p. 100032, 10 2021.

[14] P. Tiwari and V. Singh, "Diabetes disease prediction using significant attribute selection and classification approach," *Journal of Physics: Conference Series*, vol. 1714, p. 012013, jan 2021.

[15] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia computer science*, vol. 132, pp. 1578–1585, 2018.

[16] S. K. Bhoi *et al.*, "Prediction of diabetes in females of pima indian heritage: a complete supervised learning approach," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 10, pp. 3074–3084, 2021.

[17] A. A. A. Jarullah, "Decision tree discovery for the diagnosis of type II diabetes," *2011 International Conference on Innovations in Information Technology*, pp. 303–307, 2011.

[18] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463–484, 2012.

[19] M. Peng, X. Xing, T. Gui, X. Huang, Y.-G. Jiang, K. Ding, and Z. Chen, "Trainable undersampling for class-imbalance learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 4707–4714, 07 2019.

[20] M. Hammad, M. H. Alkinani, B. Gupta, and A. A. Abd El-Latif, "Myocardial infarction detection based on deep neural network on imbalanced data," *Multimedia Systems*, pp. 1–13, 2021.

[21] A. B. Hassanat, S. Mnasri, M. A. Aseeri, K. Alhazmi, O. Cheikhrouhou, G. Altarawneh, M. Alrashidi, A. S. Tarawneh, K. S. Almohammadi, and H. Almoamari, "A simulation model for forecasting COVID-19 pandemic spread: Analytical results based on the current saudi COVID-19 data," *Sustainability*, vol. 13, no. 9, 2021.

[22] A. Gosain and S. Sardana, "Handling class imbalance problem using oversampling techniques: A review," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 79–85, 2017.

[23] B. Lliu and G. Tsoumakas, "Dealing with class imbalance in classifier chains via random undersampling," *Knowledge-Based Systems*, vol. 192, p. 105292, 12 2019.

[24] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, "Learning from imbalanced data sets," in *Cambridge International Law Journal*, 2018.

[25] A. B. Hassanat, A. S. Tarawneh, G. A. Altarawneh, and A. Almuhaimeed, "Stop oversampling for class imbalance learning: A critical review," 2022.

[26] P. Vuttipittayamongkol and E. Elyan, "Improved overlap-based undersampling for imbalanced dataset classification with application to epilepsy and parkinson's disease.," *International journal of neural systems*, vol. 30, 2020.

[27] A. Jovi, K. Brki, and N. Bogunovi, "A review of feature selection methods with applications," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1200–1205, 2015.

[28] H. M. Deberneh and I. Kim, "Prediction of type 2 diabetes based on machine learning algorithm," *International Journal of Environmental Research and Public Health*, vol. 18, p. 3317, 03 2021.

[29] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, p. 11571182, mar 2003.

[30] J. Tang, S. Alelyani, and H. Liu, *Feature selection for classification: A review*, pp. 37–64. CRC Press, 2014.

[31] S. Mishra, P. Chaudhury, B. K. Mishra, and H. K. Tripathy, "An implementation of feature ranking using machine learning techniques for diabetes disease prediction," in *Proceedings of the second international conference on information and communication technology for competitive strategies*, pp. 1–3, 2016.

[32] Y. Chen and Y. Zhong, "Improved filter method for feature selection," *IOP Conference Series: Materials Science and Engineering*, vol. 569, p. 052008, 08 2019.

[33] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction," *ArXiv*, vol. abs/2206.03239, 2022.

[34] D. J. Sheskin, *Handbook of parametric and nonparametric statistical procedures*. crc Press, 2020.

[35] S. Maldonado and R. Weber, "A wrapper method for feature selection using support vector machines," *Information Sciences*, vol. 179, no. 13, pp. 2208–2217, 2009.

[36] D. Bajer, M. Dudjak, and B. Zorić, "Wrapper-based feature selection: how important is the wrapped classifier?," in *2020 International Conference on Smart Systems and Technologies (SST)*, pp. 97–105, IEEE, 2020.

[37] L. Bai, Z. Wang, Y. Shao, and N.-Y. Deng, "A novel feature selection method for twin support vector machine," *Knowledge-Based Systems*, vol. 59, 03 2014.

[38] T. Mathew and A. S, "A logistic regression based hybrid model for breast cancer classification," *Indian Journal of Computer Science and Engineering*, vol. 11, pp. 899–906, 12 2020.

[39] K. Yan and D. Zhang, "Feature selection and analysis on correlated gas sensor data with recursive feature elimination," *Sensors and Actuators B: Chemical*, vol. 212, 06 2015.

[40] P. Misra and A. S. Yadav, "Improving the classification accuracy using recursive feature elimination with cross-validation," *Int. J. Emerg. Technol*, vol. 11, no. 3, pp. 659–665, 2020.

[41] S. Sivaranjani, S. Ananya, J. Aravinth, and R. Karthika, "Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction," in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1, pp. 141–146, IEEE, 2021.

[42] J. Vallejos and S. McKinnon, "Logistic regression and neural network classification of seismic records," *International Journal of Rock Mechanics and Mining Sciences*, vol. 62, pp. 86–95, 2013.
[43] T. Ayer, J. Chhatwal, O. Alagoz, C. E. Kahn Jr, R. W. Woods, and E. S. Burnside, "Comparison of logistic regression and artificial neural network models in breast cancer risk estimation," *Radiographics*, vol. 30, no. 1, pp. 13–22, 2010.
[44] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 10 2001.
[45] S. A. Pattekari and A. Parveen, "Prediction system for heart disease using naïve bayes," *International Journal of Advanced Computer and Mathematical Sciences*, vol. 3, no. 3, pp. 290–294, 2012.
[46] G. Subbalakshmi, K. Ramesh, and M. Rao, "Decision support in heart disease prediction system using naive bayes," *Ind. J. Comput. Sci. Eng. (IJCSE)*, vol. 2, pp. 170–176, 04 2011.
[47] S. Taheri and M. Mammadov, "Learning the naive bayes classifier with optimization models," *Int. J. Appl. Math. Comput. Sci.*, vol. 23, no. 4, p. 787795, 2013.
[48] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, 2018.
[49] X. Yu, Z. Zhang, L. Wu, W. Pang, H. Chen, Z. Yu, and B. Li, "Deep ensemble learning for human action recognition in still images," *Complexity*, vol. 2020, pp. 1–23, 01 2020.
[50] K. Akyol, E. Uar, . ATLA, and M. Ucar, "An ensemble approach for classification of tympanic membrane conditions using soft voting classifier," *SSRN Electronic Journal*, 01 2022.
[51] G. Rawal, S. Yadav, and A. Singh, "Glycosylated hemoglobin (HbA1C): A brief overview for clinicians," *Indian Journal of Immunology and Respiratory Medicine*, vol. 1, pp. 33–36, 07 2016.
[52] H. Bays, R. Chapman, and S. Grandy, "The relationship of body mass index to diabetes mellitus, hypertension and dyslipidemia: Comparison of data from two national surveys," *International journal of clinical practice*, vol. 61, pp. 737–47, 05 2007.
[53] S. Gupta and S. Bansal, "Does a rise in BMI cause an increased risk of diabetes?: Evidence from india," *PLOS ONE*, vol. 15, p. e0229716, 04 2020.
[54] K. Suastika, P. Dwipayana, M. S. Semadi, and R. T. Kuswardhani, "Age is an important risk factor for type 2 diabetes mellitus and cardiovascular diseases," in *Glucose Tolerance*, ch. 5, IntechOpen, 2012.
[55] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 40–46, 2021.
[56] N. Tigga and S. Garg, "Prediction of type 2 diabetes using machine learning classification methods," *Procedia Computer Science*, vol. 167, pp. 706–716, 01 2020.