Unbalanced Learning for Early Automatic Diagnosis of Diabetes Based on Enhanced Resampling Technique and Stacking Classifier

Nawel Zemmal, Nacer Eddine Benzebouchi, Nabiha Azizi, Didier Schwab, Brahim Belhaouari

Unbalanced Learning for Early Automatic Diagnosis of Diabetes Based on Enhanced Resampling Technique and Stacking Classifier

Nawel Zemmal[a,d], Nacer Eddine Benzebouchi[a*], Nabiha Azizi[a], Didier Schwab[b], Samir Brahim Belhaouari[c]

[a]*Computer Science Department, Labged Laboratory, Badji Mokhtar Annaba University, Annaba, Algeria*
[b]*LIG-GETALP, Univ. Grenoble Alpes, Grenoble, France*
[c]*College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar*
[d]*Mathematics and Computer Science Department, Mohamed Cherif Messaadia University, Souk-Ahras, Algeria*

Corresponding author a* nasrobenz@hotmail.fr

Abstract

Diabetes is characterized by an abnormally enhanced concentration of glucose in the blood serum. It has a damaging impact on several noble body systems, mainly on the cardiovascular, renal, and visual systems. Automated screening allows early diagnosis of certain illness (such as diabetes), which generally increases the chances for successful treatment. Today, machine learning has developed considerably in the domain of medical diagnosis, especially with regard to diabetes diagnosis, and as such, thanks to the integration of the concept of unbalanced learning, which considerably reduces the generation of erroneous classification results. This general concept is dealt with from two different perspectives, i.e. at the data level through modification/balancing of the learning data set as well as at the algorithm level. The present paper takes a hybrid approach towards imbalanced learning in proposing an enhanced multimodal meta-learning method called *IRESAMPLE+St* to distinguish between normal and diabetic patients. This approach relies on the Stacking paradigm by utilizing the complementarity that may exist between classifiers. In the same focus of this study, a modified RESAMPLE-based technique referred to as *IRESAMPLE+* and the *SMOTE* method is integrated as a preliminary resampling step to overcome and resolve the problem of unbalanced data. The imbalanced Pima Indian Diabetes (*PID*) data set is optimized through the proposed *IRESAMPLE+* method, successfully operating as both an oversampling and undersampling technique, thereby reinforcing the diagnostic accuracy established by the Stacking classifier. The suggested *IRESAMPLE+St* provides a computerized diabetes diagnostic system with impressive results, A*ccuracy of 99.87%, Sensitivity of 100%, Specificity of 99.70% and AUROC of 99.90%*, comparing them to the principal related studies. The over-performing results reflect the design and engineering successes achieved with the *IRESAMPLE+St* system for the classification of diabetes.

Keywords: Unbalanced learning, multimodal classification, ensemble learning, meta-learning, diabetes diagnosis

Introduction

Diabetes is considered one of the major diseases and greatest challenges facing health systems. Due to the modern lifestyle, the incidence of diabetes in the world is in constant increasing, particularly among children. In 2015, diabetes, known as a silent killer, was the direct cause of 1.6 million deaths, and in 2012, hyperglycemia caused an additional 2.2 million deaths. More than 400 million persons in the world are living with diabetes. The number of cases of this chronic disease has quadrupled since 1990, from 108 million in 1980 to 422 million in 2014. The World Health Organization (WHO) warns that if the current tendency persists, its prevalence will increase.[1]

Diabetes is a chronic metabolic illness that cannot be cured but can be treated and controlled. It is caused by a lack/default of use of a hormone called insulin. Insulin is produced by the pancreas; it allows glucose (sugar) to enter the body's cells to be used as an energy source.

In a non-diabetic person, insulin fulfills its role well and cells have the energy they need to function. In the case of insulin insufficiency (insulinopenia) or when it does not effectively perform its function, as is the case in a person with diabetes, glucose cannot be used as fuel for cells. It then accumulates in the blood and causes an increase in the sugar level (hyperglycemia). In the long term, hyperglycemia causes certain complications affecting many noble systems of the body, such as ocular, renal, nerve, heart, and blood vessel complications.

Automated screening allows an early diagnosis of certain illness (Lamari et al., 2021; Zemmal et al., 2019), before the appearance of symptoms/complications, as well as better management and a reduction in social costs. Nowadays, Machine Learning (ML) techniques are playing a fundamental role in the evolution of the domain of medical diagnosis, particularly with respect to diagnosing diabetes. The principal goal of diagnostic aid systems is to enhance diagnostic accuracy. Actually, they are used as a second opinion by physicians to obtain the final diagnosis, which may reduce human mistakes. For this reason, several studies have been suggested concerning the automated classification/diagnosis of diabetes (Barakat, 2010; Cao et al., 2020; Choudhury & Gupta, 2019; Pradhan & Bamnote, 2015; Zou et al., 2019). The majority of medical diagnoses are based on a binary decision (for example the patient is diabetic or non-diabetic), hence the interest of classification into two categories. The classifiers generally used in the classification phase are Support Vector Machine (*SVM*) (Barakat, 2010, Pradhan & Bamnote, 2015; Abdillah & Suwarno, 2016), Gaussian Process Classification (*GPC*) (Maniruzzaman et al., 2017), Random Forest (RF) (Nai-arun & Moungmai, 2015; Zou el al., 2018), and Convolutional Neural Network (*CNN*) (Li et al., 2017).

The aim of any classification problem is to find the best features and a linear/non-linear separation boundary in order to well segregate between categories. We note that each classifier has its own manner of producing the margin that divides the classes and the resulting model varies from one classifier to another. That's why the ML Community has claimed that there is not a better (single) classifier in all cases. In other words, there are no dominant classifiers (concerning the error rate) all the others for all problems; consequently, the choice of the appropriate classifier for a given problem is not obvious and remains a challenge in the domain of ML.

Likewise, unbalanced learning or learning through imbalanced data is a common problem associated with the often unbalanced medical databases and is seen as another challenge in the area of ML, in particular with supervised learning. Class-imbalance has recently appeared in many areas of applications, including disease screening (Nnamoko & Korkontzelos, 2020), rare event prediction (Li et al., 2017), and spam filtering (Ratadiya & Moorthy, 2019) , where there are more samples available for some categories than for others. In particular, in a binary classification problem, the learning data of the minority class is much less represented than those of the majority class, which disrupts the ML

[1] https://www.who.int/diabetes/global-report/en/

algorithms (Chawla et al., 2002, Japkowicz & Stephen, 2002). Figure 1 is an illustrative example of imbalanced data.

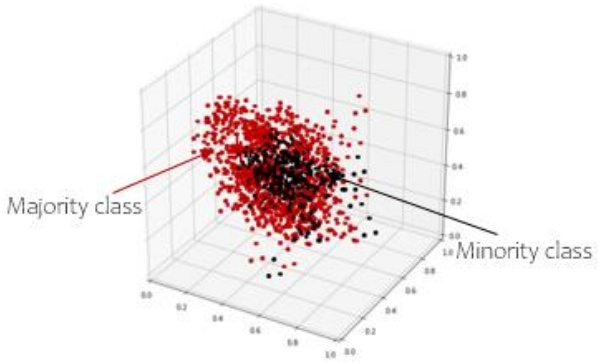

Figure 1. An example illustrating the concept of unbalanced data

In fact, all ML techniques always tend to neglect the minority class, when in most cases it is most interesting. And this is due to the fact that the separation boundary or decision boundary is converging towards the majority class, i.e. ML algorithms only learn to predict recurrent classes and thus may lead to misclassification results on the part of the classifiers, especially on the validation set.

For this purpose, it would be interesting to apply a data set balancing module so as to avoid learning errors and skewed classification. Two aspects are addressed by means of this module, that is, the processing at the data level as well as at the algorithm level. However, most approaches are addressing this issue primarily at the data level in terms of changing/balancing the learning data set through resampling techniques (*oversampling* or *undersampling*), such as the RESAMPLE and Synthetic Minority Over-sampling Technique (*SMOTE*) strategies, which are considered among the widely adopted methods to circumvent problems related to unbalanced data (Nnamoko & Korkontzelos, 2020; Li et al., 2017; Chawla et al., 2002; Japkowicz & Stephen, 2002). Oversampling is a way to rebalance the data set; it consists of increasing the number of instances belonging to the minority class by replicating them in a random way. Undersampling consists of randomly removing from the learning base instances belonging to the majority class, so as to rebalance the class distribution. Data resampling aims to prevent the problem of data imbalance to restore a more correct/normal situation, and can also be used to improve learning in order to obtain a more robust model. Relatively few methods treat the class imbalance problem at the algorithm level that consists of adapting traditional learning models so as to attenuate the bias against majority classes and to fit them to mine-related data having biased distributions (Krawczyk, 2016). Such a branch is not as widely taken up among the researchers because of its delicacy in terms of design that is directly dependent upon the dataset being used.

On the other hand, Ensemble Learning (EL) is becoming increasingly widespread as a means of dealing with class imbalance (Krawczyk, 2016). For this reason, ensemble learning (combination of classifiers) has become the widely adopted tendency in medical diagnostic support systems. Currently, EL occupies an important place in the field of Machine Learning (ML) and Artificial Intelligence (AI) because of these honorable results in various applications (Lamari et al., 2021; Abdillah & Suwarno 2016; Benzebouchi et al., 2018; Ashraf et al., 2018). The main idea of EL contributes to improving the performance of unbalanced learning by merging various classifiers so as to reduce variance, bias, and/or otherwise enhance predictions. Ensemble Learning methods in ML and AI consist of seeking to benefit from the varied expertise of different classifiers to obtain by combination a final decision model that is expected to be better than any classifier when it's considered separately. Figure 2 illustrates this process in the case of a linearly separable two-class problem. The blue dots describe first class samples and the red triangles represent the second class.

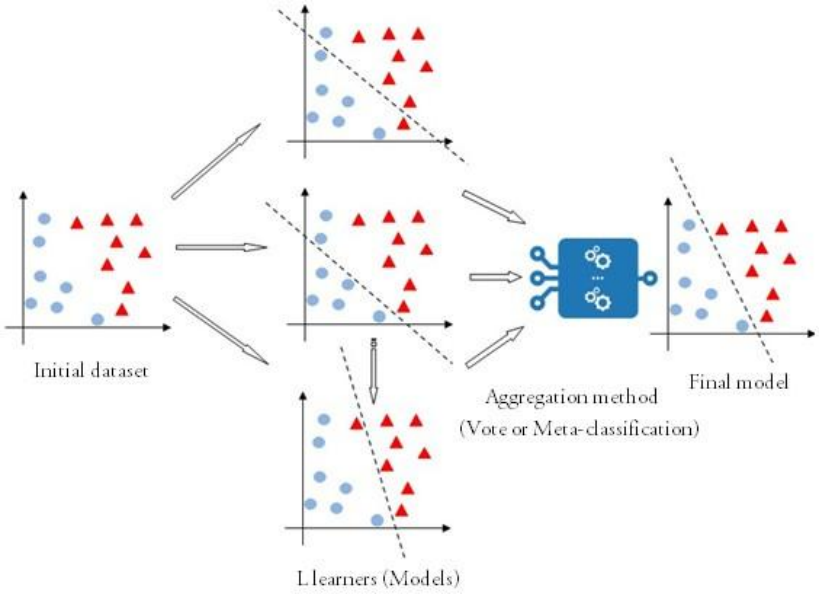

Figure 2. The general process of ensemble learning methods

The success of EL methods lies in the fact that they guarantee a lower error than its best classifier. Stacking (also called meta-classification) is an ensemble learning paradigm that allows combining various basic classifiers using a high-level meta-classifier, unlike other algorithms such as bagging and boosting based on voting methods. On the one hand, combining the characteristics of several classification models to create a new is more powerful model. The variability and diversity between different classifiers of the model are an important condition in order to take advantage of the complementarity that may exist between them. In other words, generate a set of complementary *(multimodal)* classifiers that can be combined (meta-learned) to arrive at an optimal solution. Diversity/Complementarity is considered a fundamental property of the concept of multimodality.
This article proposes a hybrid approach-based improved multimodal meta-classifier method called *IRESAMPLE+St* towards unbalanced learning for diagnosing diabetes on the basis of the Stacking algorithm by using the diversity that exists between various classifiers. With the same purpose of the study is also to obtain a balanced data set involving a data pre-processing phase through widespread resampling techniques such as the modified RESAMPLE-based method also known as *IRESAMPLE+* and the SMOTE technique, in order to form the basic classifiers of the pool as well as to get a more efficient model. The basic classifiers pool consists of five heterogeneous classifiers, namely multilayer perceptron (MLP), K-Nearest Neighbors (K-NN), Support Vector Machine (SVM), Random Forest (RF), and Naive Bayes (NB). The classification model is generated from different basic classifiers that will be used as new input data for a meta-classifier as meta-characteristics. The main task of meta-classifier or stacking aggregator, referred to as SVM, is to learn the best way to combine the different models of classification to obtain the most accurate/best classifier.

In this study, the following search questions will be exanimated closely:

- Does data preprocessing play a significant role in the performance of the proposed system?
- Which resampling technique is better adapted for balancing data?
- Which aggregation method is better suited for the combination of classifiers?
- Does the combination of classifiers have an impact on performance?

The following points highlight the principal contributions of this article:

- Hybridizing between two unbalanced data classification approaches, i.e. the data level-based and the algorithm level-based.
- Proposal for a modified resampling strategy called RESAMPLE-based *IRESAMPLE+* to enhance the classification performance on the Pima Indians Diabetes (PID) unbalanced medical data set that operates successfully as both an *oversampling* and *undersampling* technique.
- Analyzing how a meta-classification-based ensemble learning known as stacking aggregator behaves while integrating a cross-training module as well as generating a pool of complementary basic classifiers.

This article is structured as follows: an overview of the related work introduced in Section 2. Section 3 represents the meta-classification architecture as well as explains the different stages of the proposed approach. Section 4 illustrates the experimental results of this study. Finally, the conclusion of this work presented in section 5.

Related studies

Various studies have been led to develop computerized systems for the detection of diabetes. Automated diagnosis of diabetes was performed using different machine learning (ML) paradigms, some of which include techniques dealing with the unbalanced data problem and ensemble methods.

Barakat et al. (2010) proposed a diabetes diagnosis system using the SVM Classifier with RBF kernel (gamma of 0.0005 and C of 5), as well as two rules extraction techniques, SQRex-SVM and eclectic, constituting the last unit of the proposed approach, thus transforming the SVM black-box into a more comprehensible diagnostic model. An under-sampling strategy is adopted in this approach to solve the class imbalance problem, which is the K-means clustering algorithm. The results obtained showed Accuracy (Acc) of 94%, Sensitivity (Sen) of 93% and Specificity (Spe) of 94% using the leave-one-out cross-validation method.

Pradhan and Bamnote (2015) suggested a diabetes screening approach employing several data preprocessing techniques, such as normalization, discretization, and feature selection. This latter concept is applied by using the correlation-based function selector (CFS) algorithm. The SVM classifier is adopted for the classification/detection phase of diabetes using the k-fold cross-validation (CV) method with k=2. The results of the model showed that the rate of properly classified instances was 86.46% and Area Under Receiver Operating Characteristic (AUROC) curve was 83.00%.

Nai-arun and Moungmai (2015) contributed to an analytical study of the performance of different ML techniques (decision tree, artificial neural network, logistic regression, and naïve bayes) to predict diabetes, including ensemble learning methods (bagging, boosting, and random forest). A comparison of 13 classification models is carried out in order to choose the best one. Experimental results showed that the Random Forest (RF)-based ensemble method outperforms all other models with an Acc of 85.558% and an AUROC of 91.20% applying 10-fold CV.

Abdillah and Suwarno (2016) presented a tool for early diagnosis of diabetes by combining two ML techniques, namely: SVM with Radial Basis Function (RBF) kernel. This model is trained using the k-fold cross-validation technique with k=10, and its performance is assessed on the basis of performance measures derived from the confusion matrix, defined as follows: Acc = 80.22%, Sen = 82.56%, Spe = 79.12%, and AUROC = 80.84%.

Perveen et al. (2016) presented an ensemble learning approach based on the AdaBoost algorithm for diagnosing diabetes, through three age groups of the Canadian population. The analytic study showed that the performance of AdaBoost using J48 as a basic classifier outperforms the bagging using J48 as a basic classifier and the J48 decision tree algorithm as a single classifier.

Maniruzzaman et al. (2017) showed that ML models are very helpful for the classification/diagnosis of diabetes, by testing/comparing several ML algorithms, such as NB, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Gaussian Process Classification (GPC), in order to adopt the best model. Experimental results demonstrated that the GPC method gave better performance by using RBF kernel with 10-fold CV. The results indicated an Acc of 81.97%, a Sen of 91.79%, a Spe of 63.33%, a Positive Predictive Value (PPV) of 84.91% and a Negative Predictive Value (NPV) of 62.50%.

Zou et al. (2018) opted for an experimental comparison phase of several ML methods for prediction of diabetes, such as RF, J48 decision tree and Neural Network (NN) using the 5-fold CV strategy. In addition, a comparative study concerning the selection of characteristics is adopted for the reduction of dimensionality using principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR). Experimental results indicated that the RF classifier yielded better results using the mRMR technique with an Acc of 77.21%, Sen of 74.58%, Spe of 79.85%, and Matthews Correlation Coefficient (MCC) of 54.51% for the pima indians diabetes data set.

Li et al. (2017) employed K-means clustering on the basis of *CNN* and the parametric manifold learning technique with an enhanced isometric mapping algorithm (*ISOMAP*) to evaluate plantar pressure images of patients with diabetes in order to determine the important regions of plantar pressure features of the foot. The average accuracy of the clustering result was 80.00% and the manifold learning approach reached 87.20% average accuracy.

Nnamoko and Korkontzelos (2020) developed a two-stage data pretreatment approach for diabetes prediction; on the one hand, by treating outliers' instances through the interquartile range (*IQR*) algorithm. On the other hand, the *SMOTE* technique is adopted to cope with the problem of unbalanced data. In the course of this study, the authors analyzed several classifiers, namely SVM-RBF, NB, C4.5, and RIPPER, using 10-fold CV. Experimental results indicated that the C4.5 algorithm yielded the best performance with 89.50% Acc, 94.60% AUROC, 89.40% Recall, 89.50% F-score, and 83.50% Kappa.

Sarwar et al. (2020) presented an ensemble method based on voting using fifteen ML algorithms to diagnose type-2 diabetes disease. Only the four paradigms such as K-NN, SVM, ANN, and NB were considered in this study, thus representing the basic classifier pool and majority voting was adopted to reach a final decision with an accuracy of 98.60%.

Maniruzzaman et al. (2020) opted for an analysis of different classifiers such as NB, DT, RF, and Adaboost (AB) in order to maintain the most efficient classifier for the distinction between normal and diabetes patients. The logistic regression (*LR*) algorithm is considered in this work as a feature selection technique to determine diabetes risk factors according to p-value and odds ratio (OR). The RF classifier-based LR paradigm performed best with an Acc of 94.25%, a Sen of 99.57%, and an AUROC of 95.00% by applying the 10-fold CV protocol.

Ramani et al. (2020) suggested a modified *ANN* classifier approach to diabetes prediction that uses a *MapReduce* scheme in order to give a realizable frame within predictive programming paradigms for the *map* and *reduce* functions. The min-max normalization technique is performed for the pre-processing phase which consists of resizing the outputs of a value range to a new value range (e.g. 0 to 1 or - 1 to 1). The proposed study reached an accuracy level of approximately 99.60%.

Tama and Rhea (2019) presented an exploratory study concerning diabetes early detection on the basis of five various ensemble learning strategies including the bagging, random subspace, rotation forest, boosting, and DECORATE approaches. Each of them is made up of eight tree-based basic classifiers, i.e. CART, C4.5, REPT, RT, NBT, FT, BFT, and LMT. This study's performance is evaluated in terms of AUROC using a replicate cross-validation strategy ($10 \times 5CV$).

Mahabub (2019) proposed a diabetes early detection system through an analytical study of several ML paradigms like K-NN, AdaBoost (AB), DT, RF, SVM, GradientBoosting (GB), LR, MLP, MultinomialNB, ExtremeGB, GaussianNB in an aim to select three most successful classifiers (K-NN, SVM and MLP) and subsequently used to construct the base classifier pool adopting a voting approach

(hard and soft voting). On the basis of experimental results conducted as part of this work, the ensemble voting method outperforms all other separate classifiers with 86.00% Acc utilizing the 10-fold CV technique.

Rahman et al. (2020) presented a pipeline model known as "*CAMIL*" based on Multiple Instance Learning (MIL) utilizing Clustering (*UCLUST, SUMACLUST, and SWARM*) and Assembly (*SOAPdenovo2*) techniques including a Canopy cluster-based pretreatment phase for phenotype prediction of patients with type-2 diabetes through meta-genomic data (*MGWAS*). The authors applied a vocabulary-based characteristic extraction approach like the Bag of Words (*D-BoW and H-BoW*) technique as well as the *SVM-Light* classifier concerning classification. The *CAMIL H-BoW SWARM* combination showed better performance with the following results: 84.06% Acc, 85.11% F1-score, and 85.64% AUROC.

In a study analogous to Mahabub (2019), Hasan et al. (2020) examined several voting-based ensemble methods as well as different combinations of various basic classifiers (such as k-NN, DT, RF, AB, NB, and XGBoost (XB)) with the objective of selecting the most robust pool using soft weighted voting according to AUROC values in order to boost diabetes prediction. Also through this study, the authors compare MLP algorithm performance (as a unique classifier) with other suggested schemes. A preliminary step comprising outlier rejection, padding the missing or null values, data standardization (*z-score normalization*), and feature selection (*PCA, ICA, and Correlation-based approach*) is adopted within this suggested framework. Experimental results demonstrated that the pool made up of *AB & XB* basic classifiers achieved the highest performance relative to other possible combinations and MLP algorithm, resulting in an AUROC of 95.00%, a Sen of 78.90%, and a Spe of 93.40% by using the five-fold CV protocol and characteristic selection via the correlation technique.

Olisa et al. (2022) suggested a novel diabetes prediction approach based on a twice-growth deep neural network (2GDNN) classifier applied to PIMA Indian and LMCH diabetes data sets. In addition, different supervised ML algorithms are evaluated, including RF and SVM models employing repeated stratified k-fold cross-validation. A preliminary step involving feature selection and missing value imputation was performed through Spearman correlation and polynomial regression techniques, respectively. Experimental results proved that the 2GDNN model reached highest performance with a sensitivity of 97.24%, an F1-score of 97.26% using the PIMA data set and a sensitivity of 97.33% and an F1-score of 97.27% when using the LMCH diabetes data set.

Sadeghi et al. (2022) compared the performance of DNN, XGBoost and RF algorithms for the diagnosis of diabetes in TGLS cohot data. They investigate the impact of changing threshold, cost-sensitive, over and under-sampling strategies to solve the class imbalance problem. Obtained results in terms of AUROC, F1-score and G-means shows that the DNN outperforms the other algorithms. They found that the RENN under-sampling in DNN boosted ROC and Precision-Recall AUCs, g-mean and f1-measure from 85,7%, 60,3%, 71,3%, 57,5% to 86,2%, 60,8%, 77,3%, 58,3%, respectively.

A hybrid model that combines split-vote method and instance duplication was proposed in Kumar et al. (2022) to perform the prediction of diabetics in the PIMA imbalanced dataset. The concepts of over-sampling and under-sampling in conjunction with model weighting were used to increase the classification efficiency of the embedded-based ML algorithms. They obtained an accuracy of 89.32% with KNN model, 91.44% with NB model and 95.78% with SVM model.

For more information concerning computerized diabetes systems, the authors can be referred to recent reviews (Choudhury & Gupta, 2019), Sankar Ganesh & Sripriya, 2020). Most of the studies cited in the literature focus on a single classifier for the diagnostic/prediction phase of diabetes. In addition, few works use EL algorithms that are based primarily on homogeneous basic classifiers as well as voting techniques involving aggregation at the class-level, which supplies less information upon the merger. While, very few studies deal with the concept of unbalanced data at the data level through the SMOTE technique.

To the best of our knowledge, this is the first study to propose an unbalanced multimodal learning approach from an ensemble of a wide variety of heterogeneous classifiers that will be merged via a stacking aggregator or meta-classifier that handles fusion at the score-level through the rich information available at this level by applying the suggested IRESAMPLE+ technique.

Proposed IRESAMPLE+St System

The methodology of the proposed IRESAMPLE+St system is presented through a schema illustrated in Figure 3. The principal preoccupation expressed in this study consists of data balancing with the use of the enhanced RESAMPLE technique which is called *IRESAMPLE+*. Furthermore, this paper also seeks to generate an ensemble of complementary classifiers as well as the analysis of different fusion approaches using a new fusion strategy known as meta-classifier.

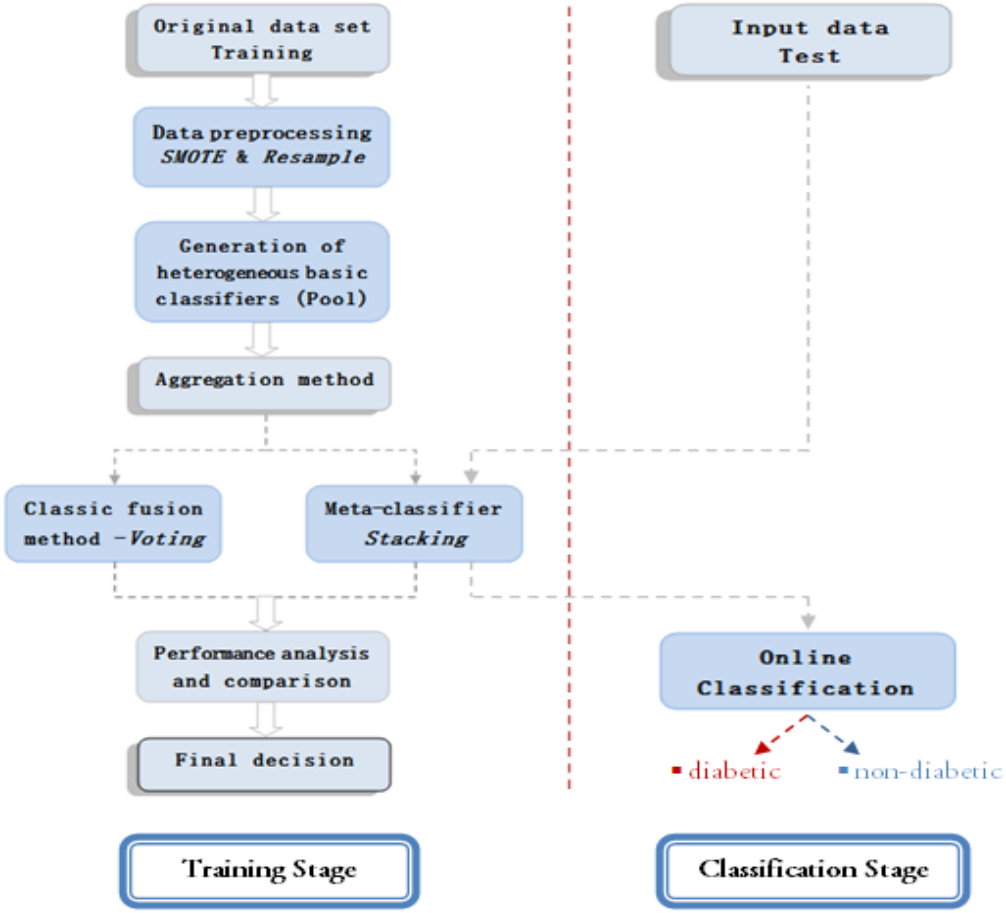

Figure 3. Proposed approach architecture for diabetes diagnosis

Data preprocessing

Data preprocessing is seen as a necessary phase in the domain of Machine Learning (ML) and data mining in order to eliminate noisy data while training classifiers on better quality data. Data resampling is a frequently used notion in an attempt to lessen the impact of data imbalance during learning. In general, this concept of resampling involves reducing the sample size of the majority class (under-sampling) or increasing the sample size of the minority class (over-sampling). Thus, for a two-class classification problem, the minority class has a relatively smaller number of samples relative to the majority class. This latter class then has a proportionally larger number of samples.

In this study, the preliminary step for data pre-processing is the make use of appropriate resampling techniques; it is about constituting a balanced data set permitting an optimal representation/distribution of the data so as to train the basic classifiers pool (promotes balanced learning). This is done via the SMOTE technique as well as a modified method based on the RESAMPLE technique, named *IRESAMPLE+*, commonly used owing to its better performance in the field of ML (Borges & Neves, 2020; Elreedy & Atiya, 2019).

The general principle of the *SMOTE* technique (supervised oversampling strategy) is to produce artificial instances to extend the boundaries of the minority class; these instances are randomly generated along line segments of a number of k-nearest neighbors that belong to the same class. Thus, this strategy makes the minority class area larger and general. The *SMOTE* strategy uses four parameters to determine the number of instances to create:

$$SMOTE \leftarrow function\ (S, P, K, C) \tag{1}$$

where,

$-S$ indicates the random number seed, $-P$ indicates the percentage of SMOTE instances to establish, $-K$ is the number of nearest neighbors to utilize, and $-C$ define the index of the nominal class value to SMOTE (minority class). Figure 4 shows an example of a synthetic instance generation.

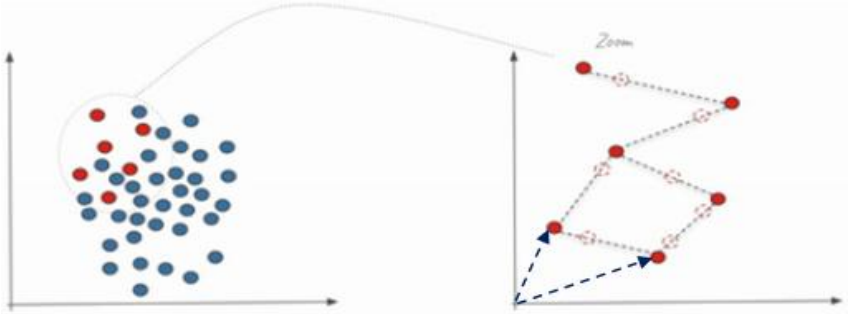

Figure 4. Illustration of the SMOTE principle

This principle is based on the difference between a minority instance and its nearest neighbor; this difference is multiplied by a random number between 0 and 1, and then added to the characteristic vector. The following pseudo-code illustrates this concept:

Pseudo-code: SMOTE (oversampling strategy)

Input: unbalanced data set (diabetes)
Output: balanced data set
Process:
1. For a sufficient number of synthetic instances do
2. Select a minority instance I
3. Select one of the nearest neighboring instances N
4. Select a random weight between 0 and 1 W
5. Create the new synthetic instance S
6. For each attribute do
7. Compute: $valueS = valueI + (valueN - ValueI) \times W$
8. End for
9. End for

The RESAMPLE technique (supervised subsampling strategy) involves creating a randomized subsample of a data set employing sampling with or without replacement, in that every instance of the sub-ensemble has the same probability of being chosen. RESAMPLE is supposed to be an impartial

representation of a cluster. The following terms refer to the number of instances to be selected as part of the RESAMPLE strategy:

$$Resample \leftarrow function\ (S, Z, B, no - replacement, V) \quad (2)$$

where,

$-S$ is the random number seed, $-Z$ indicates the size of the output data set as a percentage of the input data set, $-B$ is the bias factor to a uniform distribution of classes, $-no - replacement$ disables replacement of instances, and $-V$ reverses the choice - just available with -no-replacement-.

Figure 5 shows an example of RESAMPLE technique.

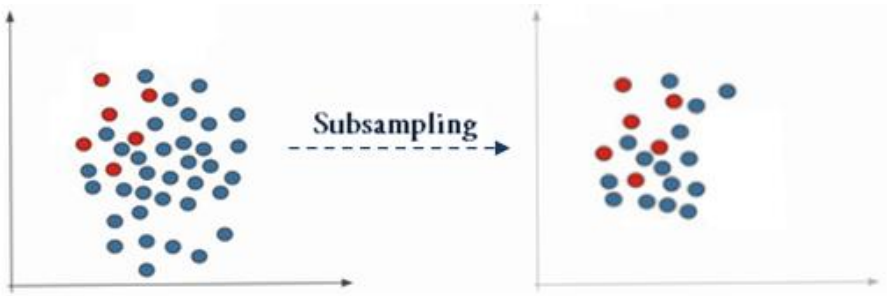

Figure 5. Illustration of the RESAMPLE principle

The goal is indeed to analyze each of these two methods (SMOTE, RESAMPLE, and IRESAMPLE+ based on the RESAMPLE approach) in order to keep the most efficient strategy.

Ensemble Classifiers approach: Basic classifiers & Aggregation strategy

Since there is no uniform classification model that can simultaneously and exhaustively solve any type of machine learning problem, it is also impossible to have an optimal classifier that can learn any distribution of learning data. This explains why it is difficult to emphasize the excellence of one ML paradigm to the detriment of another; therefore, emphasis should be placed on the use of a combination of classifiers. In effect, the aim of the combination of classifiers lies in the possibility of simultaneously constructing a set of diverse/complementary and efficient classifiers in order to achieve better decision-making through various aggregation strategies.

The *IRESAMPLE + St* multimodal learning method is in fact based on the early fusion concept (Lamari et al., 2021), which concretely means that the different outputs/characteristics issued from the basic classifiers will be combined between them "meta-features", thus constituting entries relative to a second classification model "meta-classifier" to generate the final result. Thus, no decision regarding the diagnosis of diabetes can be made once the characteristics/outputs of various classifiers are merged.

In the present case, however, the role of the basic classifiers essentially consists of different modalities, where these different classifiers bring certain types of specific information as well as their own/unique point of view. On the one hand, this is done by taking advantage of the independence between the basic classifiers, and on the other hand, by profiting from the complementarity/diversity likely to be present within the context of these classifiers. This study adopts the parallel ensemble notion that consists of the parallel production of distinct basic classifiers "m". This approach enabled us to create a set of basic classifiers following several experiments, and after a detailed review of the literature on other ML tasks. Figure 6 illustrates the proposed system structure for diabetes diagnosis.

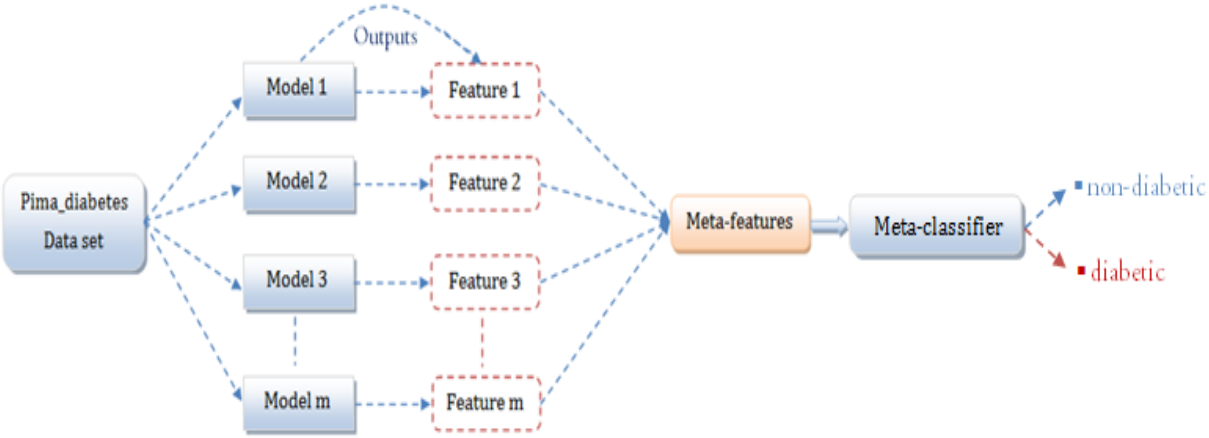

Figure 6. Proposed network architecture to predict diabetes

An appropriate aggregation strategy should be carried out after the basic learners/classifiers have been trained so as to bring together their outputs into a unique form for subsequent use in the final classification. Various strategies have been suggested allowing the combination of complementary basic classifiers of the pool, such as the linear combiner, the product combiner, and the vote combiner.

With a view to reaching a final decision regarding the proposed ensemble approach, two different class-type merging techniques (i. e. Majority Voting (MV) & Weighted Majority Voting (WMV)) are applied (Lamari et al., 2021; Benzebouchi et al., 2018), as well as the "stacking aggregator" or "meta-classifier" strategy. As for the second fusion strategy utilized during this work, namely the *meta-classifier*, it involves recourse to classifiers that, in many cases, are heterogeneous; in other words, it makes use of distinct types of learners, thus leading to a heterogeneous ensemble. The general principle of this algorithm aims to learn how to optimally aggregate the predictions of several successful ML models by means of a *stacking aggregator*.

The architecture of a Stacking scheme involves two or more basic classifiers, commonly referred to as first-level models, and a meta-model that combines the predictions of those basic models, known as the second-level model. Basic models form on the basis of an initial training dataset, and then the *meta-classifier* (or *meta-model*) is trained around the outputs/predictions of the basic models in terms of characteristics "*meta-features*".

In fact, two arguments require to be defined in the context of the construction phase of this proposed ensemble system: the basic learners "m" to be adjusted and the meta-learner responsible for setting up the aggregation phase. Those basic learners ("$m = 5$") correspond to the following algorithms, namely *Multilayer Perceptron (MLP)*, *K-Nearest Neighbors (K-NN)*, *Support Vector Machine (SVM)*, *Random Forest (RF)*, and *Naive Bayes (NB)*. At the classification/diagnosis stage, the SVM meta-learner (taking into account the characteristics of the five basic learners) is adopted because of its performance in dealing with popular problems related to binary classification. This is considered in the domain of medical diagnosis as the best binary separator.

The *IRESAMPLE+St* ensemble approach relying on the Stacking algorithm integrates the k-fold cross-training strategy with $k = 10$ (which is similar to the k-fold cross-validation technique) in order to construct the meta-model, so that all instances are taken into account to form the meta-classifier as well as the resulting model. This process consists of randomly cutting (without replacement) the original data set into k equivalent samples or k parts of approximately equal size. The first part is retained for testing and the model is trained on the $k - 1$ parts.

This means that the basic learners are then trained on the $k-1$ parts, and validated/tested by means of one of the remaining k parts so that this process is repeated k times, such that each k sub-part is utilized exactly once as a validation set. In this way, the concept of the Stacking operation "*meta-classification*" can be expressed through the following algorithm by using the k-fold cross-training method.

Algorithm: Stacking (meta-classification) with k-fold cross-training

Input: forming dataset $\mathcal{D} = \{x_i, y_i\}\ i \leftarrow 1\ to\ n$
First-level learning algorithms $C1, \ldots, CT$
Second-level meta-learning algorithm C'
Output: an ensemble of classifiers C
Process:
1. Stage1: use cross-validation technique in preparing a formation set for meta-classifier
2. Randomly split $\mathcal{D}$ into K equal-size subparts:$\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \ldots, \mathcal{D}_k\}$
3. for $k \leftarrow 1\ to\ K$ do
4. Stage 1.1: learn basic-level classifiers
5. for $t \leftarrow 1\ to\ T$ do
6. learn a classifier C_{kt} from $\mathcal{D} \backslash \mathcal{D}_k$
7. end for
8. Stage 1.2: build a forming set for meta-classifier
9. for $x_i \in \mathcal{D}_k$ do
10. $D_C = \{x'_i, y_i\}$, where $x'_i = \{C_{k1}(x_i), C_{k2}(x_i), C_{k3}(x_i), \ldots, C_{kT}(x_i)\}$
11. end for
12. end for
13. Stage2: learn a meta-classifier
14. Learn a new classifier C' based on D_C
15. Return $C(x) = C'(C_1(x), C_2(x), C_3(x), \ldots, C_T(x))$

Computational complexity

Algorithm computational complexity is an interesting study topic in the ML domain. In general, an ML model's complexity is assessed through the $Big-O$ notation. This concept can be split into two categories: *time complexity*, relating to how long the model/algorithm takes to run (which is always specified in terms relative to a certain input size "n"), and *spatial complexity*, concerning the amount of memory consumed by it.

The proposed meta-classification ensemble approach consists of a basic classifier pool "m", a number of training samples "n", and a features set "f "generated as an input for the meta-classifier by the basic classifiers. It should be mentioned that the computational cost concept is not involved in the online (real-time) classification process, it only occurs in the training phase (offline) with a low and reasonable computation time estimated at about $15\ min$.

This *IRESAMPLE+St* approach is defined as follows with respect to its computational complexity:

$$train\_time\_complexity = O(St(n^2.m.f) \text{ and } space\_complexity = O(m.f).$$

where "m" denotes the sum value with respect to the time complexity (in terms of the $Big-O$ notation) of all classifiers used (*MLP, K-NN, SVM, RF, and NB*), in which MLP is $O(n^2.f.l_i)$, K-NN is $O(n.f.k_{\text{number of neighbors}})$, SVM is $O(n_{sv}^2.f)$, RF is $O(n.log(n).f.t_{\text{number of trees}})$, and NB is $O(n.f)$. In this case, sv represents the number of support vectors, l equals the number of hidden layers, and i indicates the number of neurons in each layer.

Experimental results evaluation

With a view to evaluating the "*meta-classifier*" fusion method for diabetes diagnosis, the k-fold cross-validation technique is applied with $k = 10$. The implementation of the proposed task is performed via the *Weka* tool, an open-source library dedicated to ML algorithms, written in Java, and also giving access to well-known tools such as deeplearning4j, scikit-learn and R. The following subdivisions describe in detail the experiments carried out in the framework of this work as well as the achieved results.

Used data set

The suggested approach to diagnosing diabetes uses an ensemble of 768 instances and 8 characteristics from the Pima Indians Diabetes (PID) Database[2], of which 500 cases represent normal patients and 268 cases with diabetes. The attributes of the diabetes database are described in Table 1.

Table 1. Brief description of the attributes of the diabetes data.

Attribute name	Description	Value		
		Min	Max	Mean
Pregnancies	Number of times pregnant	0	17	3.845
Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	0	199	120.895
BloodPressure	Diastolic blood pressure ($mm\,Hg$)	0	122	69.105
SkinThickness	Triceps skin fold thickness (mm)	0	99	20.536
Insulin	2-Hour serum insulin ($mu\,U/ml$)	0	846	79.799
BMI	Body mass index ($weight\ in\ kg/(height\ in\ m)$^2)	0	67.1	31.993
DiabetesPedigreeFunction	Diabetes pedigree function	0.08	2.42	0.472
Age	Age (years)	21	81	33.241
Outcome	Class variable (negative or positive)	0	1	-

Evaluation criteria

In order to assess the performance of the proposed model, we use performance measures commonly used in the medical domain which are given as follows in Eqs. (6), (7), (8), (9), (10), (11), and (12):

$$Sensitivity\ (SEN) = \frac{TP}{TP + FN} \tag{6}$$

$$Specificity\ (SPE) = \frac{TN}{TN + FP} \tag{7}$$

$$Positive\ Predictive\ Value\ (PPV) = \frac{TP}{TP + FP} \tag{8}$$

$$Negative\ Predictive\ Value\ (NPV) = \frac{TN}{TN + FN} \tag{9}$$

$$Accuracy\ (ACC) = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$F - measure\ (F - score)\ = 2\,.\ \frac{PPV\ .\ SEN}{PPV + SEN} \tag{11}$$

[2] https://www.kaggle.com/

$$Matthews\ Correlation\ Coefficient\ (MCC) = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (12)$$

Where, True Positives (TP), True Negatives (TN), False Positives (FP), and False negatives (FN) are described in Table 2.

Table 2. Binary classification confusion matrix (diabetes)

	Positive test (P)	Negative test (N)
Patients with diabetes (P)	TP	FN
Non-diabetic patients (N)	FP	TN

The suggested method was also evaluated according to the area under receiver operating characteristic (AUROC) curve (Bradley (1997)).

Results and Discussion

Several experiments concerning the constitution of the basic classifier pool and, more precisely, on the number of selected classifiers are carried out during this study (Azizi & Farah, 2012; Zemmal et al., 2016), by adopting five distinct models which are the following: the Multilayer Perceptron (*MLP*), the K-Nearest Neighbors (*K-NN*), the Support Vector Machine (*SVM*), the Random Forest (*RF*), and the Naive Bayes (*NB*). Such classifiers are moreover formed from the original data set, i.e. without any prior processing, so as to be able to better evaluate and compare the results obtained through these experiments.

1 Experimental set-up

The set-up parameters as well as each algorithm's performance are described as follows:
As for the *MLP* classifier, ideal outcomes are attained through adjustment of the following settings: "$hidden\_layers: 'a' = ((attribs + classes)/2) = 5$", "$learning\_rate = 0.3$", "$momentum = 0.2$", "$number\ of\ epochs = 500$", and "$validation\_threshold = 20$". The relevant performances achieved are 75.39% ACC, 60.80% SEN, 83.20% SPE, 63.29% F-1 score, 44.92% MCC and 79.30% AUROC.
Concerning the *K-NN* classifier, the settings on the basis of K-number of neighbors to use, $nearest\_Neighbour\_Search\_Algorithm$, $distance\_Weighting$, and $window\_Size$ are set so as to obtain optimized results by using "$K = 2$", "$Algorithm: Euclidean\_Distance - R\ first - last$", "$distance\_Weighting: No$", and "$window\_Size = 0$". The corresponding values obtained are 72.66% ACC, 55.20% SEN, 82.00% SPE, 58.49% F-1 score, 38.37% MCC and 74.20% AUROC.
The optimal results are reached with respect to *SVM* classifier by setting up the following main variables: $kernel$, $gamma$, and C, where "$kernel: Polynomial$" with a $gamma$ value of 0.5 and parameter C equal to 3. The correspondent results reached are 77.47% ACC, 55.60% SEN, 89.20% SPE, 63.27% F-1 score, 48.42% MCC and 72.40% AUROC.
The tuning settings with their optimized outcomes obtained using the *RF* classifier are determined as follows: "$maxDepth: unlimited$", "$numFeatures = (< 0 = int\ (log\_2\ (\#predictors) + 1))$", and "$numTrees = 100$". The correspondent results achieved include 74.87% ACC, 59.00% SEN, 83.40% SPE, 62.10% F-1 score, 43.51% MCC and 81.50% AUROC.
The NB classifier performs best by setting both $displayModelInOldFormat$ and $useKernelEstimator$ parameters to $false$ and $useSupervisedDiscretization$ parameter to $true$. The correspondent performances represent with an ACC of 76.30%, a SEN of 61.20%, a SPE of 84.40%, an F-1 score of 64.30%, a MCC of 46.78% and an AUROC of 81.90%.

The experiments were conducted with a Windows-7 operating system under the following hardware setup: Intel® Core™ i7-8750H CPU @ 2.20 GHz up to 4.10 GHz processor with RAM 64 GB and DDR4-2666, LPDDR3-2133 Memory.

2 Balancing of the data set using oversampling and subsampling strategy

The present study examines the way in which the pool of basic classifiers can be used as part of the pre-processed/balanced data of the PID database. In other words, this suggested method is based on results obtained through data balancing techniques (*SMOTE* and the modified RESAMPLE-based method called *IRESAMPLE+*). As an illustration of this recourse to resampling techniques concerning the proposed approach, the different phases to follow to balance the data using the *SMOTE* filter and the suggested *IRESAMPLE* + method are thus presented below. It should be noted that Figure 7 shows that there is a large non-uniformity in data distribution across the classes of the data set (diabetes_PID) employed in this work.

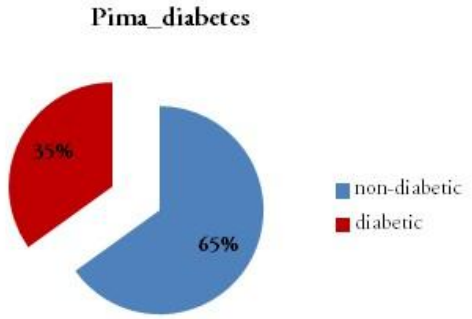

Figure 7. Class imbalance of the diabetes data set

The disparity in final classes such as tested_negative $_{\text{non-diabetic}}$ and tested_positive $_{\text{diabetic}}$ can result in incorrect diagnostic results. Thus, the classifier may be more biased towards the high concentration classes (non-diabetic) than the low concentration class (diabetic), which may also lead to the inaccurate classification of patients to detect the resulting category.

Taking into consideration the nature of the unbalance of the output class, namely tested_positive (diabetic class), an oversampling strategy, *SMOTE*, is applied to the diabetes data set. The supervised filtering strategy is applied using the settings of function (1) as follows:

$$SMOTE \leftarrow function\ (1, 86.6, 5, 2) \tag{1}$$

The make use of the SMOTE method (in this configuration) permitted to increase the size of the minority class (diabetic) from 268 to 500 in order to dispose of a balanced database and to obtain significant results. Figure 8 presents a comparative diagram across the raw data set vs. the application of the SMOTE technique.

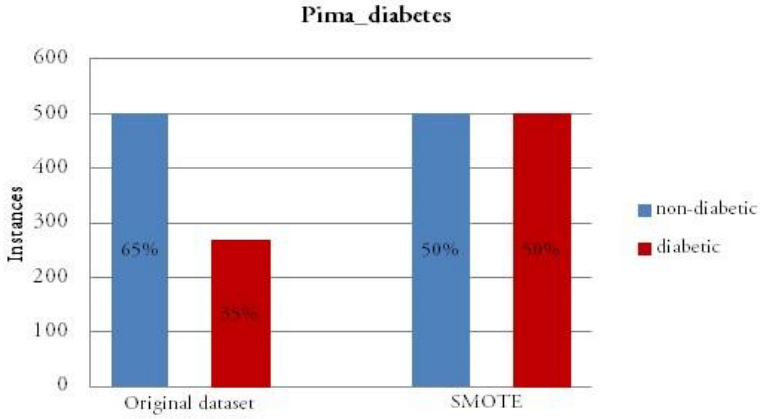

Figure 8. Applying the SMOTE filter to class distribution

On the other hand, the subsampling strategy, *RESAMPLE*, is also practiced on the majority class (non-diabetic) to reduce its size from 500 to 268, so that both classes possess an identical number of instances. This supervised filtering strategy (*RESAMPLE*) can be configured in the following way:

$$Resample \leftarrow function\ (1, 69.8, 1, no - replacement, F) \qquad (2_1)$$

Such a setup has in effect engendered a great loss of data (instances), leading ultimately to unsatisfactory results. The present study aims at modifying/improving the general functioning and principle of this technique so that it can operate at the same time as an oversampling and undersampling strategy; this is achieved through an appropriate adjustment at the level of equation (2) parameters, while also introducing an iterative structure enabling a significant balancing of the data. The following pseudo-code highlights this aspect:

Pseudo-code: *IRESAMPLE+* Procedure (oversampling and subsampling strategy)

Input: Original unbalanced data set (diabetes)
Output: more efficient balanced data set
Process:
1. Begin
2. Initialisation
2. Repeat
3. $Resample \leftarrow function\ (1, 100, 0, no - replacement, F)$ (2_2)
4. Until (number of class $1_{\text{non-diabetic}}$ instances = number of class 2_{diabetic} instances)
5. End

Meanwhile, the use of the enhanced RESAMPLE strategy succeeded, on the one hand, in increasing the size of the minority class (diabetics) from 268 to 384 and, on the other hand, in decreasing the size of the majority class (non-diabetics) from 500 to 384. The purpose of this strategy is to provide a balanced data set with exactly the same number of instances in the two classes while preserving the maximum of instances, with the objective of minimizing bias and thus reinforcing the classification/diagnosis accuracy established by the classifier. Figure 9 graphically illustrates the application of the *RESAMPLE* technique and the modified RESAMPLE-based method known as *IRESAMPLE+*.

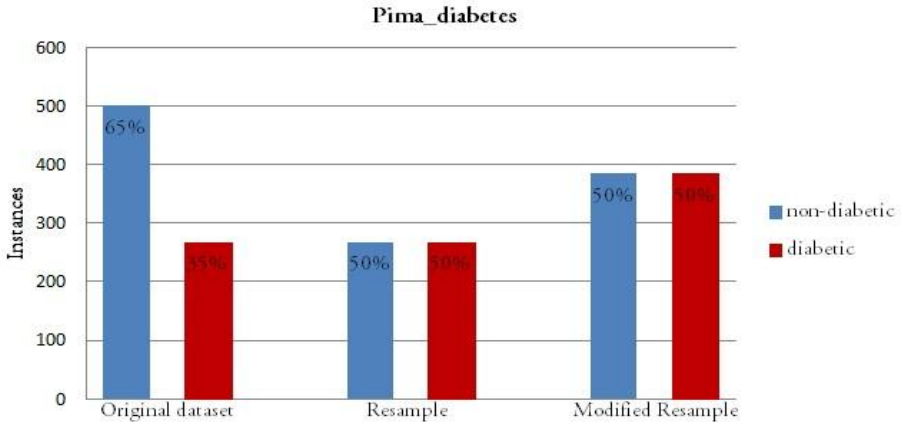

Figure 9. Practice both RESAMPLE and IRESAMPLE+ methods for class distribution

Indeed, the use of the SMOTE (oversampling) method permitted to augment the minority class size by randomly adding artificial instances, which can lead to the overlearning problem. The idea of the RESAMPLE (subsampling) technique consists of the random removal of the majority class instances, which can eventually lead to the suppression of the relevant information. In order to overcome the different problems mentioned above, the present study suggests a new strategy called IRESAMPLE+ as a compromise between the two previously discussed techniques (SMOTE and RESAMPLE), that can operate at the same time as an oversampling and undersampling strategy enabling a meaningful analysis and relevant balancing of the PID data set. The effect of these different resampling strategies regarding the distribution of the classes is illustrated in Figure 10.

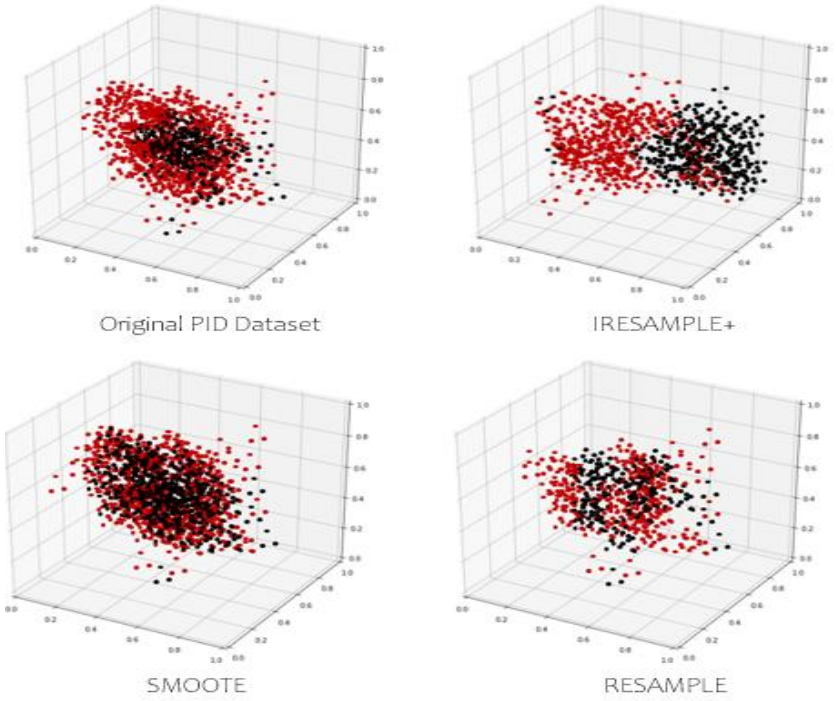

Figure 10. Comparison of different resampling techniques on the distribution of PID classes-diabetics (black circles) and non-diabetics (dark red circles)-

3 Models performance analysis

This article aims to separately analyze the specific performance of each classifier performed, whether before or after the application of data resampling techniques (SMOTE and IRESAMPLE+) and this, according to the performance factors mentioned above: ACC, SEN, SPE, PPV, NPV, F-score, MCC,

and AUROC. The following tables (Table 3 and Table 4) summarize the obtained outcomes from five models (MLP, K-NN, SVM, RF, and NB).

Table 3. Obtained results using the basic classifiers separately before resampling the diabetes data set

Classification method	Performance measures (%)							
	ACC	SEN	SPE	PPV	NPV	F-score	MCC	AUROC
MLP	75.39	60.80	83.20	66.00	79.80	63.29	44.92	79.30
K-NN	72.66	55.20	82.00	62.20	77.40	58.49	38.37	74.20
SVM	77.47	55.60	89.20	73.40	78.90	63.27	48.42	72.40
RF	74.87	59.00	83.40	65.60	79.10	62.10	43.51	81.50
NB	76.30	61.20	84.40	67.80	80.20	64.30	46.78	81.90

Table 4. Obtained results using the basic classifiers separately after resampling the diabetes data set

Filtre	Classification method	Performance measures (%)							
		ACC	SEN	SPE	PPV	NPV	F-score	MCC	AUROC
SMOTE	**MLP**	77.40	79.80	75.00	76.10	78.80	77.90	54.86	83.30
	K-NN	79.20	84.60	73.80	76.40	82.70	80.30	58.74	83.00
	SVM	79.60	83.40	75.80	77.50	82.00	80.30	59.37	79.60
	RF	80.80	83.80	77.80	79.10	82.80	81.40	61.71	87.80
	NB	76.50	80.80	72.20	74.40	79.00	77.50	53.20	84.60
IRESAMPLE+	**MLP**	98.31	99.00	97.70	97.70	98.90	98.30	96.62	97.30
	K-NN	98.83	97.90	98.70	98.70	98.00	97.80	97.67	98.90
	SVM	99.37	99.00	99.70	99.70	99.00	99.60	99.22	99.60
	RF	98.96	98.40	98.50	98.50	98.50	98.00	98.22	98.90
	NB	95.44	97.70	93.20	93.50	97.50	95.50	90.98	98.30

On the basis of findings described in Tables 3 and 4, it is clear that, after the application of resampling techniques, the classifiers' performance is still superior to that observed in the original data. In addition, the proposed *IRESAMPLE+* filter provides better results compared to the SMOTE filter, and therefore the performance of the classifiers is considerably optimized. Figure 11 more visibly illustrates this comparative study using only AUROC values.

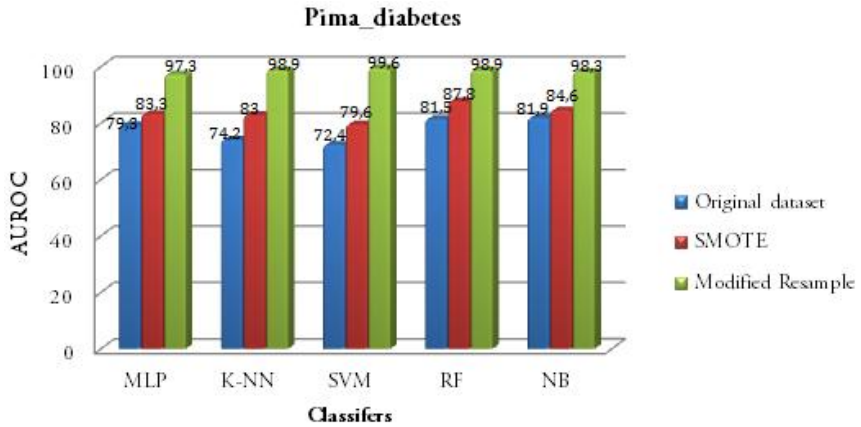

Figure 11. Impact of applying SMOTE and IRESAMPLE+ filters on the performance of classifiers

In order to enhance the sensitivity and specificity of the proposed multimodal learning as well as the global performance in terms of diagnosis, on the one hand; moreover, on the other hand, to reach a final decision about this study, three aggregation approaches are applied: the Majority voting (*MV*), the Weighted Majority Voting (*WMV*), and the SVM meta-classifier (or *stacking aggregator*). Thus, the obtained results by using the *SMOTE, RESAMPLE* and *IRESAMPLE+* techniques are indicated in Table 5.

Table 5. Obtained results from three aggregation paradigms used for diabetes data set

Filtre	Aggregation method	Performance measures (%)							
		ACC	SEN	SPE	PPV	NPV	F-score	MCC	AUROC
SMOTE	**MV**	81.50	85.40	77.60	79.20	84.20	82.20	63.19	81.50
	WMV	**82.41**	**84.70**	**79.80**	**80.40**	**84.10**	**82.10**	**65.43**	**82.30**
	SVM meta-classifier	80.50	84.60	78.40	79.30	81.80	80.90	61.00	80.50
RESAMPLE	**MV**	76.68	75.00	78.40	77.60	75.80	76.70	53.39	76.70
	WMV	**78.45**	**77.60**	**80.30**	**79.40**	**77.20**	**78.50**	**55.92**	**78.00**
	SVM meta-classifier	72.02	69.00	75.00	73.40	70.80	72.00	44.11	72.00
IRESAMPLE+	**MV**	98.83	99.70	97.90	98.00	99.70	98.80	97.67	98.80
	WMV	99.72	100	99.50	99.50	100	99.70	99.46	99.90
	SVM meta-classifier	**99.87**	**100**	**99.70**	**99.70**	**100**	**99.90**	**99.74**	**99.90**

Based on the performance described above (Table 4 and Table 5), we observe that the concept of Ensemble Learning (*EL*) using *Resampling* techniques (*SMOTE*, *RESAMPLE*, and *IRESAMPLE+*) is clearly preferable over separate use of classifiers. In addition, it should be noted that the *WMV* method generated a more accurate classification with the use of the SMOTE & RESAMPLE filters, while the SVM meta-classifier aggregation paradigm using the *IRESAMPLE+* strategy is much more efficient and robust than the other solutions investigated in this work. This approach has also shown greater reliability compared to current findings using the same data set (PID). Figure 12 illustrates the ROC curves of the three aggregation paradigms employed (MV, WMV, and Stacking_SVM) using the proposed *IRESAMPLE+* filter.

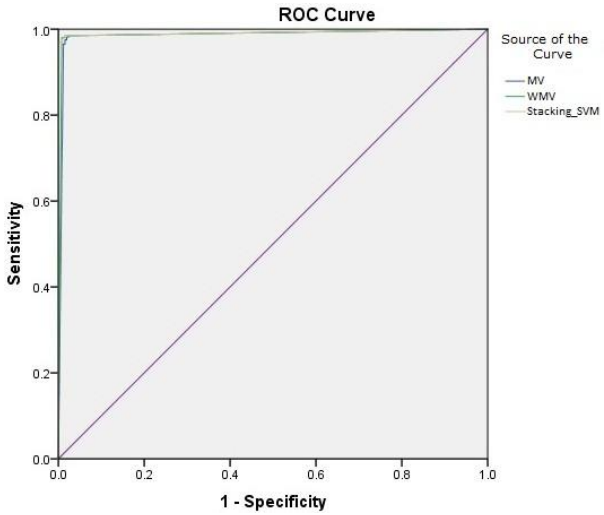

Figure 12. ROC curve of three aggregation paradigms used

By noting that all curve points for the three techniques used are positioned in the upper half of the ROC space, resulting in a good ROC curve, especially for the SVM meta-classifier model.

The following diagrams demonstrate that the meta-classification approach considerably reduces the error rate compared to separate learning of classifiers and gives better performance using the ACC (Figure 13), SEN (Figure 14), and SPE (Figure 15) evaluation criteria.

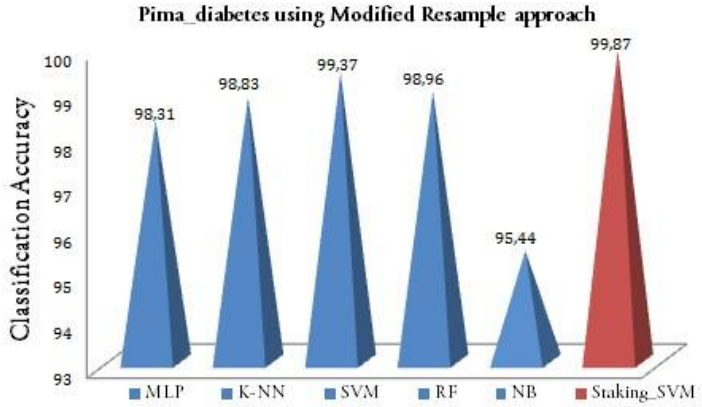

Figure 13. Comparison of Accuracy rate between basic classifiers and the meta-classification method

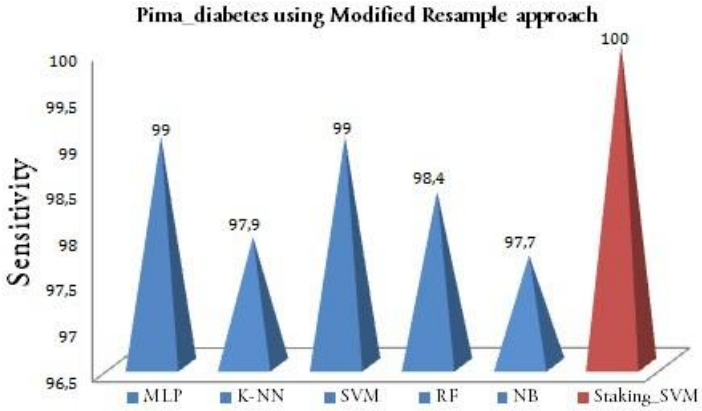

Figure 14. Comparison of Sensitivity rate between basic classifiers and the meta-classification paradigm

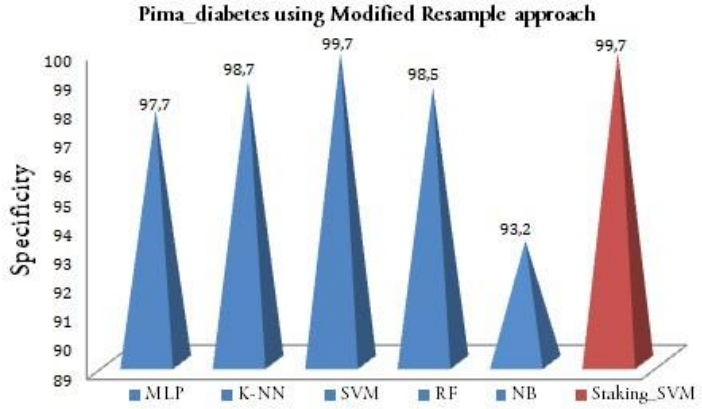

Figure 15. Comparison of Specificity rate between basic classifiers and the Stacking approach

It should be pointed out as well that individual classifiers "m" (such as MLP, K-NN, SVM, RF, and NB) are referred to as the first-level learners (basic classifiers) while the meta-learner is identified as the second-level learner. For each of these first-level classifiers, the outputs are generated on the basis of the original training data set "p", while a new data set "f" is created to form the second-level meta-classifier. The first-level classifier predictions made are given as input characteristics to the meta-classifier or stacking aggregator with the same class labels as in the source data set. The resulting meta-features will, therefore, have a training set of "$m \cdot f$" size.

An experimental study concerning the *SVM meta-classifier* performance using different kernel methods such as the *polynomial* kernel, the *Radial Basis Function (RBF)* kernel and the *PUK* kernel with multiple

K numbers (5 and 10) of the cross-validation approach accompanied by Table 6 illustrating the obtained findings.

Table 6. Obtained results with various kernel optimizations for two K-fold protocols

k	Used kernel	Performance measures (%)							
		ACC	SEN	SPE	PPV	NPV	F-score	MCC	AUROC
5	Polynomial	99.73	100	99.50	99.50	100	99.70	99.46	99.70
	RBF	**99.87**	**100**	**99.70**	**99.70**	**100**	**99.90**	**99.74**	**99.90**
	PUK	99.73	99.50	100	100	99.50	99.70	99.46	99.70
10	Polynomial	99.61	99.50	99.70	99.70	99.50	99.60	98.42	99.60
	RBF	**99.87**	**100**	**99.70**	**99.70**	**100**	**99.90**	**99.74**	**99.90**
	PUK	99.61	99.50	99.70	99.70	99.50	99.60	98.42	99.60

Comparing the obtained results (Table 6), we find that the SVM meta-classifier using the Radial Basis Function (RBF) kernel gave the highest performance and this, regardless of the K-fold value (5-times/10-times) of the cross-validation approach. Resulting in 99.87% Accuracy (ACC) with Sensitivity (SEN) 100%, Specificity (SPE) 99.70%, Positive Predictive Value (PPV) 100%, Negative Predictive Value (NPV) 99.50%, F-score 99.90%, Matthews Correlation Coefficient (MCC) 99.74%, and AUROC 99.90%, which leads to the improvement of the diagnostic performance. The key parameters to tune for the SVM meta-classifier are specified as follows: RBF kernel with a $gamma$ value of 0.5 and the C parameter equal to 3. Table 7 presents other important measures of performance evaluation of the suggested method provided by the *Weka* tool, such as the Kappa statistic (classification reliability), the Mean Absolute Error (MAE), the Root Mean Square Error (RMSE), the Relative Absolute Error (RAE), and the Root Relative Square Error (RRSE), which are defined as follows in Eqs. (13), (14), (15), (16) and (17):

$$Kappa = \frac{P_o - P_e}{1 - P_e} \tag{13}$$

where, P_o indicates the relative observed accordance between the classification raters, and P_e is the hypothetical probability of chance accordance.

$$MAE = \frac{\sum_{i=1}^{n} |P_i - O_i|}{n} \tag{14}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (P_i - O_i)^2}{n}} \tag{15}$$

$$RAE = \frac{\sum_{i=1}^{n} |P_i - O_i|}{\sum_{i=1}^{n} |\bar{O} - O_i|} \tag{16}$$

$$RRSE = \frac{\sum_{i=1}^{n} (P_i - O_i)^2}{\sum_{i=1}^{n} (\bar{O} - O_i)^2} \tag{17}$$

where, P_i represent the predicted target, O_i represent the observed target, and n is the number of observations.

Table 7. illustration of the Kappa statistic, MAE, RMSE, RAE and RRSE rate of the proposed method

Correctly Classified Instances	767	99.8698 %
Incorrectly Classified Instances	1	0.1302 %
Kappa statistic	0.9974	
Mean absolute error	0.0013	
Root mean squared error	0.0361	
Relative absolute error	0.2604 %	
Root relative squared error	7.2168 %	
Total Number of Instances	768	

Table 8 shows the confusion matrix for this study, which measures the quality of the proposed classification/diagnostic system. This table indicates how many correct predictions for each class and the number of incorrect predictions for each class organized according to the predicted class. Each row corresponds to an actual class while each column corresponds to a predicted class.

Table 8. Confusion Matrix of the suggested approach

	Positive test (P)	Negative test (N)
Patients with diabetes (P)	384	0
Non-diabetic patients (N)	1	383

As shown in Tables 7 and 8, 384 instances of diabetes were correctly classified as diabetes cases by the proposed method, 383 healthy instances are properly classified as non-diabetes cases. In summary, 767 instances are accurately labeled and only one is not, resulting in 99.74% Cohen's kappa coefficient (k) with MAE = 0.13%, RMSE = 3.61%, RAE = 0.2604%, and RRSE = 7.2168%, which further proves the reliability of the suggested methodology.

The analysis of the experimental results led to the conclusion that the combination of classifiers is preferable to the separate learning of classifiers (considerably reduces the error rate), and that the combination of multimodal features using the meta-classification method has demonstrated its effectiveness in diagnosing diabetes. The objective of this study is access to the conception of a computer-aided diagnosis (CAD) system for diabetes disease which surpasses other systems of literature; this thanks to the aggregate of the characteristics/outputs of several classifiers in meta-classification architecture as well as to the IRESAMPLE+ technique.

Table 9 exemplifies the key approaches proposed in the literature and introduces a comparative study of the suggested method performance against the leading existing approaches of diagnosing diabetes in the literature. This table is organized in terms of accuracy (ACC) measurement.

Table 9. Performances comparison of the proposed method against other different ML algorithms available in the literature

Authors	Used dataset	Data preprocessing method	Classification method	ACC (%)	SEN (%)	SPE (%)	AUROC (%)
(Hasan et al., 2020)	Pima Indians Diabetes (PID)	Outlier rejection, filling the null values, *z-score normalization*, and *Correlation-based* feature selection	Ensemble method AdaBoost+XGBoost (AB & XB) using soft weighted voting	-	78.90	93.40	95.00
(Varma et al., 2014)	PID	Elimination of missing values (EMV)	Modified Gini index-Gaussian fuzzy decision tree	75.80	-	-	-
(Bozkurt et al., 2014)	PID	-	NN-Distributed Time Delay Networks (DTDN)	76.00	53.33	88.75	-
(Singh & Singh, 2020)	PID	-	Stacking with SMO	79.00	78.90	-	73.20
(Choubey & Paul, 2016)	PID	Feature selection (FS) using the Genetic Algorithm (GA)	Genetic Algorithm (GA)-MLP NN	79.13	79.10	-	84.20
(Iyer et al., 2015)	PID	Transformation and Feature selection	Naive Bayes (NB)	79.57	-	-	-
(Abdillah & Suwarno, 2016)	PID	-	SVM with Radial Basis Function (RBF) kernel	80.22	82.56	79.12	80.84
(Nai-arun & Moungmai, 2015)	Sawanpracharak Regional Hospital (SRH)	Transformation and Selection	Random Forest (RF)	85.56	-	-	91.20
(Mahabub, 2019)	PID	Normalization	Ensemble voting method (K-NN, SVM and MLP)	86.00	-	-	-
(Ramezani et al., 2018)	PID	Imputation and OT linear dimension reduction algorithm	Logistic Adaptive Network-based Fuzzy Inference System (LANFIS)	88.05	92.15	81.63	-
(Alghamdi et al., 2017)	Henry Ford FIT Hospitals in metropolitan Detroit in U.S	Discretization, FS (MLR & Entropy), and the SMOTE filter	Ensemble method with voting technique	89.00	99.70	74.70	92.20
(Nnamoko & Korkontzelos, 2020)	PID	IQR and SMOTE techniques	C4.5	89.50	89.40	-	94.60
(Chen & Pan, 2018)	Hospital of WenZhou Medical Univ	Deletion of records is not in numerical format	LogitBoost	89.63	-	-	96.30
(Maniruzzaman et al., 2017)	PID	Normalization using the median technique	Gaussian Process Classification with RBF kernel	91.97	91.79	63.33	-
(Nilashi et al., 2017)	PID	Self-Organizing Map (SOM) + PCA	Neural Network (NN)	92.28	-	-	-
(Maniruzzaman et al., 2020)	National Health and Nutrition Examination Survey (NHANES)	Feature selection using LR algorithm	Random Forest (RF)	94.25	99.57		95.00
(Nai-Arun & Sittidech, 2014)	SRH	FS using the Gain Ratio Algorithm	Bagging	95.31	-	-	-
(Yilmaz et al., 2014)	PID	Modified K-means Algorithm	Modified K-Means + SVM	96.71	97.31	95.06	-
(Sarwar et al., 2020)	database created	-	Ensemble model (K-NN, SVM, ANN, and NB) using majority voting	98.60	-	-	-
Olisa et al. (2022)	PID and LMCH diabetes	FS and EMV using Spearman correlation and polynomial regression techniques	a twice-growth deep neural network (2GDNN)	97.25	97.24	-	-

Suggested IRESAMPLE+St	PID	Data balancing through the proposed IRESAMPLE+ strategy	Stacking aggregator-based Multi-modal learning using the SVM meta-classifier	99.87	100	99.70	99.90

As observed in Table 9 with respect to performances of related studies available in the literature, it indicates as though all systems operate differently in terms of data pre-processing technique and classification method applied towards diagnosing diabetes using the same data set, i.e. PID. For instance, Nilashi et al. (2017) used the SMO+PCA technique as a preprocessing method and the NN paradigm for the classification stage. Bozkurt et al. (2014) investigated different neural networks (NN)-based classifiers, i.e. distributed time delay networks (DTDN), anticipation networks, learning vector quantization, cascade networks, probabilistic neural networks, and time delay networks using the original data set. Also, Singh and Singh (2020) made use of the original data set, i.e. without any pre-processing by adopting the stacked generalization approach on the basis of different kernels of the SVM classifier such as Linear-SVM, Polynomial-SVM, RBF-SVM and Sigmoid by using the meta-learner SMO. Iyer et al. (2015) employed data pretreatment methods such as transformation and feature selection (FS) in order to apply the NB classifier to the resulting set. Choubey and Paul (2016) opted for the feature selection (FS) method as the data set pre-processing utilizing the genetic algorithm (GA) with the MLP classifier to perform the diagnostic phase. Yilmaz et al. (2014) proposed a modified K-means algorithm for the pre-processing phase and to be used with the SVM classifier. Ramezani et al. (2018) reported a logistic adaptive network-based fuzzy inference system (LANFIS) utilizing imputation and OT linear dimension reduction algorithm with regard to the data pretreatment. Varma et al. (2014) removed missing values from the data set employing a modified gini index-gaussian fuzzy decision tree as a classification approach.

It should be noted that the majority of studies do not consider the notion of unbalanced data (especially at the data level) in the diabetes classification, which represents a major problem with regard to the field of ML, leading to erroneous classification results. However, there is relatively limited research (Nnamoko & Korkontzelos, 2020; Alghamdi et al., 2017) that has taken into account such an unbalanced data concept in adopting the *SMOTE* technique. Likewise, most of the approaches mentioned in the literature as approaches to diabetes diagnosis are often based or on a standard approach relying on a single classifier's point of view or on a vote-based ensemble approach (class-level aggregation) that provides less information (available) at the time of fusion.

The proposed IRESAMPLE+St approach by the present study addresses with the unbalanced data concept at both grades, i.e. at the data level as well as at the algorithm level. The former concept is processed through the suggested enhanced RESAMPLE method called *IRESAMPLE+* (*so as to equilibrate distributions and/or eliminate difficult samples*); the second is performed by adopting a stacking aggregator (*which includes the k-fold CV technique*)-based multimodal ensemble approach.
By comparing the performances summarized in Table 8 and with the studies reported previously in the related work section, the proposed IRESAMPLE+St approach clearly surpasses all the state-of-the-art approaches by obtaining the most optimal results in terms of accuracy, sensitivity, specificity, AUROC, and by Cohen's kappa. This confirms as well that the suggested IRESAMPLE+St meta-classification method offers a better and more precise early diagnosis of diabetes disease with 99.87% ACC, 100% SEN, 99.70% SPE, 99.90% AUROC, and 99.74% Cohen's kappa.

With the aim of the proposed model generalization along with further experimental evaluation, different benchmark datasets having unbalanced data are utilized/tested. Pima Indians Diabetes, Parkinson's and

Table 10. Comparative results of the proposed approach using other well-known medical data sets (#I: instances, #P: positive instances, #N: negative instances, #ACC: accuracy)

Approach / *Datasets*	#I	Before Pre-processing			After Pre-processing		
		#P	#N	#ACC(%)	#P	#N	#ACC(%)
Diabetes	768	268	500	75.65	384	384	99.87
Diabetic Retinopathy Debrecen	1151	611	540	79.30	575	575	95.48
Breast Cancer	286	85	201	67.48	143	143	82.87
Unbalanced dataset	856	12	844	98.59	428	428	100
Parkinson's	195	147	48	91.70	97	97	95.45
Cardiac Catheterization Diagnostic	3504	2372	1132	92.40	1752	1752	97.38
Prostate Cancer Dataset	506	213	293	94.78	253	253	98.70
VA Lung Cancer	138	40	124	92.32	82	82	96.90

Cardiac Catheterization Diagnostic are taken from UCI Machine Learning Repository. Prostate Cancer and VA Lung Cancer datasets from Vanderbilt Biostatistics Wiki (Frank & Harrell, 2016).

Table 10 presents the results achieved in terms of accuracy (ACC) and data balancing by the proposed approach and this, before and after the application of the *IRESAMPLE+* suggested resampling technique on each medical dataset used.

Based on Table 10, the experiments conducted on the various medical datasets validates the meaningful performance of the *IRESAMPLE+St* proposed approach.

Conclusion

Diabetes is considered one of the gravest diseases in the world threatening human health, also called the silent killer. Screening allows early diagnosis of certain diseases, before the appearance of symptoms/complications, as well as better management and a decrease of the social cost.

The main objective of this paper is to design a robust medical decision support system, more precisely diabetes disease by analyzing several advanced paradigms of the combination of classifiers taking into account the context of the often unbalanced medical bases. This approach uses the meta-classification paradigm that has shown better performance with the addition of the enhanced resampling module known as *IRESAMPLE+*, which makes our system more robust and reliable. The results demonstrated that ensemble learning with the preprocessed data provides a very low error rate.

However, the limitation of the suggested approach is that only the diabetes database is treated/tested and that will be overcome in subsequent work by applying other unbalanced data sets through analysis of other advanced resampling techniques. In addition, there are also plans to build a new data set that includes modern healthcare to predict diabetes comprises a urine test, and the glycated hemoglobin A1c (HbA1c) test.

In conclusion, the suggested IRESAMPLE+St approach delivers a more robust Clinical Decision Aid System (CDAS) that permits diabetologists to rapidly diagnose patients with diabetes at an early stage while also providing a second opinion to the doctors with high accuracy in order to support their therapeutic decisions.

References

Abdillah, A., & Suwarno, S. (2016). Diagnosis of Diabetes using Support Vector Machines with Radial Basis Function Kernels. International Journal Of Technology, 7(5), 849-858

Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., & Sakr, S. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. PLOS ONE, 12(7): e0179805.

Ashraf, M., Zaman, M., & Ahmed, M. (2018). Using Ensemble StackingC Method and Base Classifiers to Ameliorate Prediction Accuracy of Pedagogical Data. Procedia Computer Science, 132, 1021–1040.

Azizi, N., & Farah, N. (2012). From static to dynamic ensemble of classifiers selection: Application to Arabic handwritten recognition. KES Journal, 16(4), 279-288.

Barakat, N., Bradley, A. P., & Barakat, M. N. H. (2010). Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus. IEEE Transactions on Information Technology in Biomedicine, 14(4), 1114–1120.

Benzebouchi, N. E., Azizi, N., Aldwairi, M., & Farah, N. (2018, April). Multi-classifier system for authorship verification task using word embeddings. 2018 2nd IEEE International Conference on Natural Language and Speech Processing (ICNLSP) (pp. 1–6). Algiers, Algeria: IEEE.

Borges, T. A., & Neves, R. F. (2020). Ensemble of machine learning algorithms for cryptocurrency investment with different data resampling methods. Applied Soft Computing, 106187. doi:10.1016/j.asoc.2020.106187

Bozkurt, M.R , Yurtay, N., Yilmaz, Z., & Sertkaya, C. (2014). Comparison of different methods for determining diabetes. Turk J Elec Eng & Comp Sci, 22(4):1044–1055.

Bradley, A.P.: The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recogn. 30(7) (1997) 1145–1159

Cao, L., Dey, N., Ashour, A.S., Fong, S., Sherratt, R. S., Wu, L., & Shi, F. (2020). Diabetic plantar pressure analysis using image fusion. Multimed Tools Appl, 79, 11213–11236

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

Chen, P., & Pan, C. (2018). Diabetes classification model based on boosting algorithms. BMC Bioinformatics, 19(1), 109.

Choubey, D.K., Paul, S. (2016). GA_MLP NN: A Hybrid Intelligent System for Diabetes Disease Diagnosis. Int J Intell Syst and Appl, 8(1), 49-59

Choudhury, A., & Gupta, D. (2019). A Survey on Medical Diagnosis of Diabetes Using Machine Learning Techniques. In: Kalita J., Balas V., Borah S., Pradhan R. (Eds.), Recent Developments in Machine Learning and Data Analytics. Advances in Intelligent Systems and Computing, vol 740. Springer, Singapore.

Elreedy, D., & Atiya, A. F. (2019). A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for Handling Class Imbalance. Information Sciences, 505, 32-64.

Frank, E., & Harrell, J. R. (2016). Vanderbilt Biostatistics Wiki. [http://biostat.mc.vanderbilt.edu/DataSets]. Department of Biostatistics, University of Vanderbilt.

Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. IEEE Access, 8, 6516 - 76531

Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques, arXiv:1502.03774v1

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent Data Analysis, 6(5), 429–449.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. Prog Artif Intell, 5, 221–232.

Kumar, M.S., Khan, M., Rajendran, S., Noor, A., Dass, S., & Jayagopal, P. (2022). Imbalanced Classification in Diabetics Using Ensembled Machine Learning. Computers, Materials and Continua. 72. 10.32604/cmc.2022.025865

Lamari, M. . (2021). SMOTE–ENN-Based Data Sampling and Improved Dynamic Ensemble Selection for Imbalanced Medical Data Classification. In: Saeed, F., Al-Hadhrami, T., Mohammed, F., Mohammed, E. (eds) Advances on Smart and Soft Computing. Advances in Intelligent Systems and Computing, 1188. Springer, Singapore. https://doi.org/10.1007/978-981-15-6048-4_4

Li, J., Fong, S., Hu, S., Chu, V. W., Wong, R. K., Mohammed, S., & Dey, N. (2017). Rare Event Prediction Using Similarity Majority Under-Sampling Technique. In: Mohamed A., Berry M., Yap B. (eds) Soft Computing in Data Science. SCDS 2017. Communications in Computer and Information Science, vol 788. Springer, Singapore

Li, Z., Dey, N., Ashour, A., Cao, L., Wang, Y., Wang, D., McCauley, P., Balas, V., Shi, K., & Shi, F. (2017). Convolutional Neural Network Based Clustering and Manifold Learning Method for Diabetic Plantar Pressure Imaging Dataset. Journal of Medical Imaging and Health Informatics, 7, 639-652

Mahabub, A. (2019). A robust voting approach for diabetes prediction using traditional machine learning techniques. SN Appl. Sci., 1(12), 1667

Maniruzzaman, M., Kumar, N., Menhazul Abedin, M., Shaykhul Islam, M., Suri, H. S., El-Baz, A. S., & Suri, J. S. (2017). Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. Computer Methods and Programs in Biomedicine, 152, 23–34.
Maniruzzaman, M., Rahman, M. J., Ahammed, B., & Abedin, M. M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. Health Inf Sci Syst, 8(1), 1-14
Nai-arun, N., & Moungmai, R. (2015). Comparison of Classifiers for the Risk of Diabetes Prediction. Procedia Computer Science, 69, 132–142.
Nai-Arun, N., & Sittidech, P. (2014). Ensemble Learning Model for Diabetes Classification. Advanced Materials Research, 931-932, 1427–1431.
Nilashi, M., Ibrahim, O., Dalvi, M., Ahmadi, H., & Shahmoradi, L. (2017). Accuracy Improvement for Diabetes Disease Classification: A Case on a Public Medical Dataset. Fuzzy Information and Engineering, 9(3), 345–357.
Nnamoko, N., & Korkontzelos, I. (2020). Efficient Treatment of Outliers and Class Imbalance for Diabetes Prediction. Artificial Intelligence in Medicine, 104, 101815
Olisah, C.C., Smith, L., & Smith, M. (2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. Computer Methods and Programs in Biomedicine, 220, 106773
Perveen, S., Shahbaz, M., Guergachi, A., & Keshavjee, K. (2016). Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. Procedia Computer Science, 82, 115–121.
Pradhan M., & Bamnote G.R. (2015). Efficient Binary Classifier for Prediction of Diabetes Using Data Preprocessing and Support Vector Machine. In: Satapathy S., Biswal B., Udgata S., Mandal J. (eds) Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014. Advances in Intelligent Systems and Computing, vol 327. Springer, Cham
Rahman, M. A., LaPierre, N., & Rangwala, H. (2020). Phenotype Prediction from Metagenomic Data Using Clustering and Assembly with Multiple Instance Learning (CAMIL). IEEE/ACM Transactions on Computational Biology and Bioinformatics, 17(3), 828-840
Ramani, R., Vimala Devi, K. & Ruba Soundar, K. (2020). MapReduce-based big data framework using modified artificial neural network classifier for diabetic chronic disease prediction. Soft Comput. In press
Ramezani, R., Maadi, M., & Khatami, S. M. (2018). A novel hybrid intelligent system with missing value imputation for diabetes diagnosis. Alexandria Engineering Journal, 57(3), 1883-1891
Ratadiya, P., & Moorthy, R. (2019). Spam filtering on forums: A synthetic oversampling based approach for imbalanced data classification. ArXiv, abs/1909.04826
Sadeghi, S., Khalili, D., Ramezankhani, A., Mansournia, M. A., & Parsaeian, M. (2022). Diabetes mellitus risk prediction in the presence of class imbalance using flexible machine learning methods. BMC medical informatics and decision making, 22(1), 36. https://doi.org/10.1186/s12911-022-01775-z
Sankar Ganesh P.V., and Sripriya P. (2020). A Comparative Review of Prediction Methods for Pima Indians Diabetes Dataset. In: Smys S., Tavares J., Balas V., Iliyasu A. (eds) Computational Vision and Bio-Inspired Computing. ICCVBIC 2019. Advances in Intelligent Systems and Computing, vol 1108. Springer, Cham
Sarwar, A., Ali, M., Manhas, J. & Sharma, V. (2020). Diagnosis of diabetes type-II using hybrid machine learning based ensemble model. Int. j. inf. tecnol. 12, 419–428
Singh N., Singh P. (2020). A Stacked Generalization Approach for Diagnosis and Prediction of Type 2 Diabetes Mellitus. In: Behera H., Nayak J., Naik B., Pelusi D. (eds) Computational Intelligence in Data Mining. Advances in Intelligent Systems and Computing, vol 990. Springer, Singapore
Tama, B.A., & Rhee, K. (2019). Tree-based classifier ensembles for early detection method of diabetes: an exploratory study. Artif Intell Rev, 51, 355–370
Varma, K.V.S.R.P., Rao, A.A., Sita Maha Lakshmi, T., & Nageswara Rao, P.V. (2014). A computational intelligence approach for a better diagnosis of diabetic patients. Computers & Electrical Engineering, 40(5), 1758–1765.
Yilmaz, N., Inan, O., & Uzer, M. S. (2014). A New Data Preparation Method Based on Clustering Algorithms for Diagnosis Systems of Heart and Diabetes Diseases. J Med Syst, 38:48(5).
Zemmal, N., Azizi, N., Ziani, A., Benzebouchi, N.E., and Aldwairi, M. "An Enhanced Feature Selection Approach based on Mutual Information for Breast Cancer Diagnosis," 2019 6th International Conference on Image and Signal Processing and their Applications (ISPA), 2019, pp. 1-6, doi: 10.1109/ISPA48434.2019.8966803
Zemmal, N., Azizi, N., Dey, N., & Sellami, M. (2016). Adaptative S3VM semi supervised learning with features cooperation for breast cancer classification. Journal of Medical Imaging and Health Informatics, 2016, 6(4), 957–967
Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques. Frontiers in Genetics, 9, 1-10.