STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **13**, April 2025, pp 1595–1609.
Published online in International Academic Press (www.IAPress.org)

A Metaheuristic for Fuzzy Density Based SVM and Confidence SMOTE for Early Prediction of Diabetes

Asma Driouich [1], Abdellatif El Ouissari [2,*], Karim El Moutaouakil [3] and Ismail Akharraz [1]

[1] *Engineering, Mathematicals and Informatiques laboratory, Faculty of Sciences, uiz, Agadir, Morocco*
[2] *LaR2A Laboratory, Faculty of Sciences, Abdelmalek Essaadi University, Tetouan, Morocco*
[3] *Laboratory of Mathematics and data science , Faculty Polydisciplinary of Taza, USMBA*

Abstract Early detection of diabetes, based on observable features, plays a crucial role in preventing serious complications in diabetic patients. In this study, we propose a classification model called SMOTE Density Based Fuzzy Support Vector Machine (SMOTE-DB-FSVM), based on FSVM, to better detect diabetes. Our approach is based on five main steps: data cleaning, density-based filtering, feature selection to identify the most important attributes, calculation of a confidence score for each point in the minority class, and use of SMOTE to balance the data. In addition, we compare different versions of the kernel functions in the SVM model to optimize classification results, using metaheuristics to estimate the parameters of these kernels. The proposed SMOTE-DB-FSVM algorithm has been evaluated in diabetes datasets, including the PIMA diabetes database, and the results show a clear improvement in the early detection of diabetes with this method.

Keywords DB-Support vector machine, Class Imbalance, Classification, Fuzzy logic, Diabet, Artificial Intelligence, Machine learning

DOI: 10.19139/soic-2310-5070-1348

1. Introduction

The prediction of diabetes has been the focus of much artificial intelligence research in recent years. Indeed, machine learning has proved highly effective in detecting diabetes, which will facilitate doctors' tasks and speed up patient treatment. Artificial intelligence tools thus play a crucial role in helping to control, to some extent, the exponential growth of the diabetes phenomenon. In addition, there is a growing number of machine learning models for prediction and classification, and these models greatly simplify the understanding of data [1, 2, 3]. For example, SVM, XGBoost, DT, RNN, RF, and others [4, 5, 6]. Some also choose to use two models at the same time to study the phenomenon, e.g. CNN-LSTM [7], hybrid FSRF model[8], hybrid CNN-SVM model [9].

However, when it comes to medical data, we are often confronted with unbalanced data, which makes interpretation difficult for simple machine learning models. This is why data rebalancing techniques are often used. One of the best-known and most effective methods for balancing data is SMOTE, and its various versions. Thousands of articles have been published in this field. Some are specifically related to SMOTE, while others concern hybrid models. These works often demonstrate the efficacy of this method for treating diabetes, although there are still differences in accuracy between the models and the versions developed. What interests us in this article is research into diabetes prediction using data rebalancing algorithms. Of course, oversampling presents the problem of generating noise or erroneous points, which is one of the challenges of SMOTE. A large body of research has proposed solutions to this problem, such as K-means SMOTE [10], LR-SMOTE [11] and others.

*Correspondence to: Abdellatif Elouissari (Email: a.elouissari@uae.ac.ma). Department of Mathematics, Faculty of Sciences, Abdelmalek Essaadi University, Tetouan, Morocco.

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 2025 International Academic Press

In this paper, we propose a confident version of the density-based support vector machine for early detection of diabetes, called SOMTE Density Based Support Vector Machine (SOMTE-DB-FSVM), which proceeds in five steps: (a) data cleaning, (b) density-based filtering, (c) feature selection to identify the most important attributes, (d) calculation of a confidence score for each point in the minority class, and (e) use of SMOTE to balance the data. then we solve the dual SVM to detect support vectors. The proposed method is compared to other known classifiers on a PIMA unbalanced diabetes data set and the Germany data sets such as Naive Bayes, Decision tree, artificial neural network, Support vector machine and Density based fuzzy SVM [17, 31].
This article is organized as follows: In the second section, we provide an overview of relevant previous works and other existing imbalanced classification methods of diabetes prediction based on PIMA dataset. The third section presents our proposed approach, called SOMTE Density Based Support Vector Machine (SOMTE-DB-FSVM), in which we explain the mathematical theories behind the approach, as well as the different kernel functions and the reasons for selecting each one. The fourth section discusses the experimental results obtained with our SOMTE-DB-FSVM approach on PIMA diabetes and Germany datasets and its comparison with other known classifiers. Finally, we conclude the paper with section 6.

2. Review of different prediction models for unbalanced PIMA data

To predict diabetes by using the Indian Pima Diabetes Dataset (PIDD), many works have used machine learning (ML) methods[18],[19]. A number of closely related works are discussed in this section.

The prediction of diabetes using machine learning methods and models is making steady progress, with constant developments every day. What's more, research is also being carried out into the pre-processing and processing of data, before it is introduced into the analysis phases. Consequently, the pre-processing stage plays just as crucial a role as the prediction stage, for one simple reason: diabetes-related data is extremely sensitive. Our paper focuses on two main aspects: on the one hand, the pre-processing stage using the SMOTE method, and on the other hand, the prediction phase with the calculation of the confidence rate for each element, integrated as a constraint in the optimization problem proposed in our recent work. In order to improve the quality of our study, we have compared our results with the most recent methods in this field. It should be noted that there is a large body of work exploiting oversampling models. The following section discusses the most recent contrubitions proposed in the literature:

For research focusing on early detection we start with that of J. J. Khanam [25], who used machine learning algorithms to compare them and give the best algorithm that works very well with the PIMA dataset. He summarized that Logistic Regression (LR) and Support Vector Machine (SVM) are the top ones. They also built a neural network model as an additional task for their paper. Another paper, similar to the first, is by Tigga et. al in [26], who also demonstrated that the random forest model offers the best predictions. This study is based mainly on data highlighting two specific effects: patients' lifestyle and family history. The same model was then applied to the PIMA database, proving the originality and effectiveness of the proposed model.

One study that uses a mixture of two techniques, that proposed by Shuja et al [27], involves the construction of a classification algorithm that takes into account a data pre-processing step using the SMOTE class balancing method. Subsequently, they chose machine-learning classification methods that have been shown to be effective in diabetes prevention, SVM (Support Vector Machine), MLP (Multi-Layer Perceptron), Simple Logistic, and Decision Tree. Experimental results demonstrate the effectiveness of this pre-processing in improving prediction performance.

Delshi and all in [28] they proposed an approach that was used to diagnose Diabetes mellitus (DM). This method benefits from the Farthest First (FF) clustering algorithm and the Sequential Minimal Optimization (SMO) classifier algorithm. they used Farthset First (FF) to group the data into number of clusters and Support Vector Machine (SVM) to classify the output to diabetic and non-diabetic patients. Until now the problem of imbalanced data remains one of the complex problems of binary classification in the machine learning. The Imbalanced data problem occurs when we have two classes of different sizes, i.e. the number of data in one class (called minority) is smaller than the number of data in the other class (called majority). Several works have been done to solve this type of problem. Re-sampling of learning data can be done in two ways, eliminating data from

the majority class (under-sampling), or inflating the minority class with artificial data or duplicating existing observations (oversampling). This problem has attracted the attention of many practitioners and several methods and techniques have been proposed to develop under- and over-sampling techniques. Over-sampling techniques depend on duplicating samples or generating new samples; practitioners sometimes use pooling procedures to determine appropriate areas for synthetic sample generation. Oversampling is very simple and easy to solve the problems of imbalanced data. Random oversampling randomly duplicates minority class instances until the desired class distribution is reached. One of the most effective techniques for oversampling is the SMOTE technique, which has been proposed by Chawla[20]. This technique generates artificial data instead of reproducing existing observations. To generate artificial data in the minority class, this technique mainly depends on randomly choosing a time x_1 in this class, then selecting among the k minority class neighbors a time x_2 closest to the first chosen time, is easily a new data x generates between x_1 and x_2 by the following formula: $x = x_1 + w.(x_2 - x_1)$, where w is a random weight in $[0, 1]$. But a common disadvantage of SMOTE is that it does not distinguish between the limits of the decision and it can therefore choose a sample from the majority class with the consequence that there are disturbances on oversampling. SMOTE is a simple and efficient method, indeed several extensions and modifications have been made to develop this method. borderline-SMOTE1 and borderline-SMOTE2[21] are two techniques that focus on the minority class region, instead of randomly selecting data by SMOTE, borderline-SMOTE1 targets the data of the boundary; and by k nearest neighbours, the algorithm can determine the data if it is rejected as noise.Borderline-SMOTE2 allows the generated sample to be reconciled with the minority class, specifying the interpolation weight between 0 and 0.5. Cluster-SMOTE helps to avoid noise generation and to overcome imbalances between categories, using the k-means method to cluster the minority class and by the SMOTE method by generating samples in safe areas[22]. Nekooeimehr and Lai-Yuen[23] propose another oversampling technique called Adaptive Semi-Unsupervised Weighted Oversampling (A-SUWO) which is based on hierarchical clustering and has used clustring to increase oversampling quality. For undersampling, Lin and al[24] use the SMOTE method to group the majority class into k groups, where k is the sample number of the minority class. The method eliminates all data except the centers of the k clustring, resulting in equal size classes. However, this technique deletes data that has a value in the class. In the study by Md. Ashraf Uddin [29], they proposed a machine learning model for diabetes prediction, using various machine learning methods. To evaluate the proposed model, they carried out a pre-processing process, obviously, the cleanup with the removal of missing values, subsequently data normalization and label coding. To deal with data imbalance, they used the SMOTE technique to improve algorithm performance. In paper [30], the authors proposed a new version of SMOTE that generates new data with the guarantee of not creating noisy data, based on the notion of Fuzzy C-Means.

3. Improved SMOTE for imbalanced data

In this section, we present the proposed method and the various steps involved, including some comparisons with other artificial intelligence models. The main idea is to allow oversampling algorithms to avoid generating noise, while identifying the most reliable points. The method will be compared with several versions of SMOTE on unbalanced datasets, since our approach is based on SMOTE. In addition, the method will also be compared with versions of SVM with different kernel functions, particularly in the case of non-linearly separable data, which requires the use of these kernel functions. The proposed contrubition is based on 5 steps: data cleaning, density-based filtering, feature selection/importance, calculation of a confidence score for each minority class data point, SMOTE. Each step plays an essential role in improving the prediction of diabetic patients. Data cleansing is one of the key steps in ensuring that artificial intelligence algorithms function properly. In addition, the DB-FSVM step eliminates noisy data, as it is impossible to generate new data from noisy or anomalous data. That's why this step is one of the most important. Calculating the confidence rate is also crucial, as DB-FSVM cannot eliminate 100% of noisy data. So, to ensure that SMOTE can select a point in the minority class as the center for generating new data, confidence points are needed. This confidence rate is calculated on the basis of sound mathematical concepts. Finally, the use of SMOTE enables new data to be generated without the risk of additional noise, while balancing the two classes, thus ensuring a suitable environment for reliably training machine learning models.

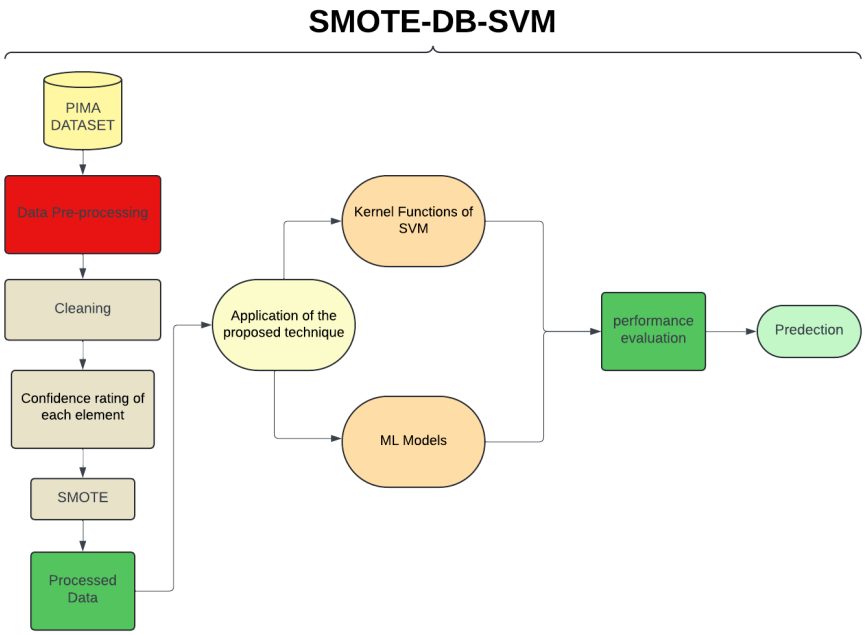


Figure 1. The proposed contrubition integrated kernel functions for PIMA diabetes prediction

3.1. The preprocessing steps of the SMOTE-DB-FSVM algorithm

In this section, we take a detailed look at the steps involved in the SMOTE-DB-FSVM method. The figure 2 shows the steps followed by the method for class balancing with a high level of confidence.

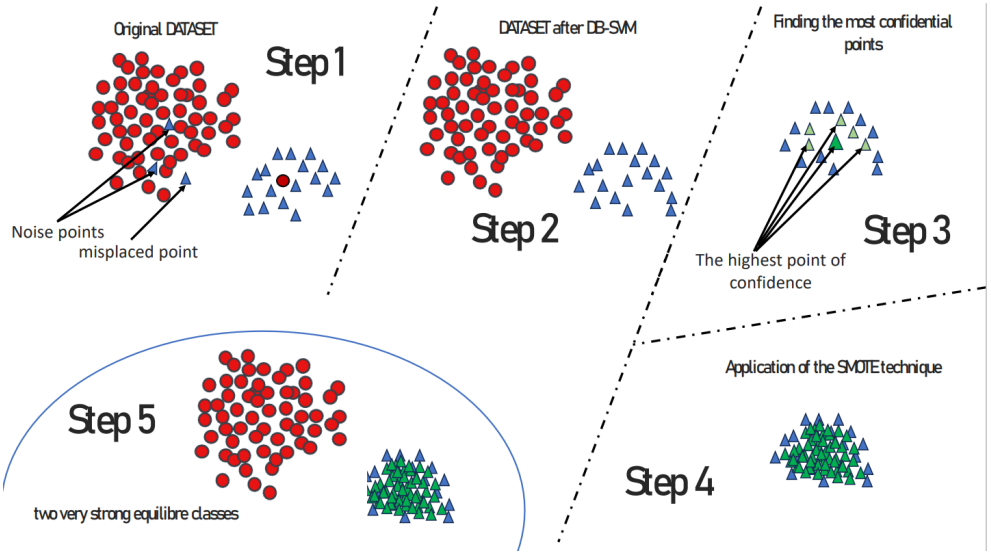


Figure 2. The preprocessing steps of the SMOTE-DB-FSVM algorithm diagram

Data cleaning: in the context of medical data, the issue of data sensitivity is crucial. On the one hand, medical data is scarce; on the other hand, University Hospitals (CHU) often refuse to share their patient databases for security reasons. In this regard, based on the PIMA and Germany datasets, we only eliminated missing data and, in our case, we cannot replace missing data with the mean or median of the observed values. For one simple reason, the diabetes dataset is sensitive, and we can't use any technique for missing data without guaranteeing the preservation of data originality. What's more, the SMOTE method will increase the size of the data, which means we'll be generating data. So, replacing it after adding more data in the preprocessing phase will run the risk of losing the originality and quality of the data. Next, we used DB-FSVM [17] to eliminate noise and anomaly points, as it has proven effective in removing this type of data.

Density based theory

Using the density based SVM technique, we can define the noise data as follows: Let BD be a set of N samples $x_1, ..., x_N$ labeled, respectively, by $y_1, ..., y_N$, distributed via k class $C_1, ..., C_k$.

For a given data set BD, a non-negative real r and an integer mp, there exist three kind of samples. A sample x is called C_i-Noise Point (NP_i) if $|C_i \cap B(x, r)| < mp$. We use this definition to determine all the existing noises in the data set and eliminate them.

By using SVM and to construct the hyperplane that separates the classes, we need only the support vectors that always exist at the bound, for this reason, we are looking for cord points and the deleted. The cord data is defined by the following criteria: A sample x is called C_i-Cord Point (CP_i) if $|C_i \cap B(x, r)| \geq mp$ and $x \in \overbrace{envol(C_i)}^{o}$

After removing the noise and string data in the original dataset, we can easily use the SMOTE algorithm to balance the data. In this regard, we proposed to use SMOTE only for the most confidence points of the minority class and to create the SVM boundary decision based on the most confidence support vectors.

3.2. Confidence Ratings and New Optimization Model for DB-FSVM

At this stage, we determine the confidence score for each point in the minority class and select the point with the highest score to serve as the center of the SMOTE algorithm, in order to generate new points.

Synthetic confidence evaluation:

To calculate the degree of confidence, we use the statistical concept of quartiles. We use quartiles to evaluate the confidence of each element. Quartiles are the variable values that divide an ordered data set into four equal groups. These three values are denoted by Q_1, Q_2, and Q_3, where : Q_2 is the median, i.e. the central value that divides the data set into two equal parts, Q_1 is the value below which no more than 25% of observations fall, and Q_3 is the value below which no more than 75% of observations fall. To use these quartiles to evaluate the confidence of each element, we identified the most important feature in the PIMA dataset. To achieve this, we employed a technique involving regression coefficients, by using Python. This technique uses linear model coefficients to determine the value that each feature contributes to the prediction. By fitting features to a linear regression model, we predict the target label. Feature importance is then determined by the absolute values of the linear model coefficients associated with each feature. The higher these absolute values, the greater the influence of the feature on the model. As shown in 3, we found that BMI is the most important feature in the given dataset. Therefore, we will use it to calculate the degree of confidence.

We denote by i the minority class and consider a synthetic sample s_j (generated by the SMOTE method). The confidence level is $DC_i(s_j)$ is defined as follows:

$$DC_i(s_j) = \min(\frac{Tot_i(s_j)}{mp}, 1)$$

where mp min point defined previously and $Tot_i(x) = |C_i \cap B(x, r)|$ is the number of minority samples in $B(x, r)$, with $r > 0$ chosen experimentally. The figure 4 shows the estimation of two types of synthetic point confidence s_j.

$\forall j$, the larger $DC_i(s_j)$ is, the more likely s_j is to be selected as a support vector (of course if it is a border simple). In our case the "large" means $DC_i(s_j) > Q_3\%$. To inform the optimization problem about this decision rule, $DC_i(s_j)$ is introduced in the constraints of the SVM dual model; see Equ.1.

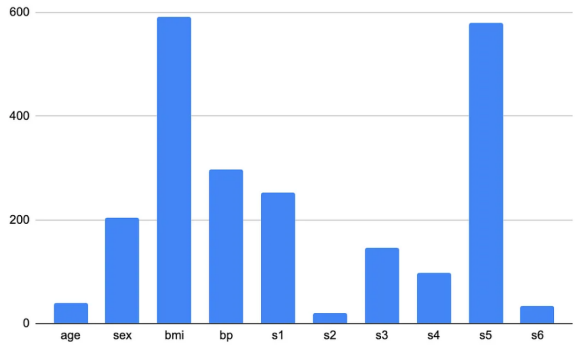


Figure 3. Feature Importance Values of PIMA Dataset

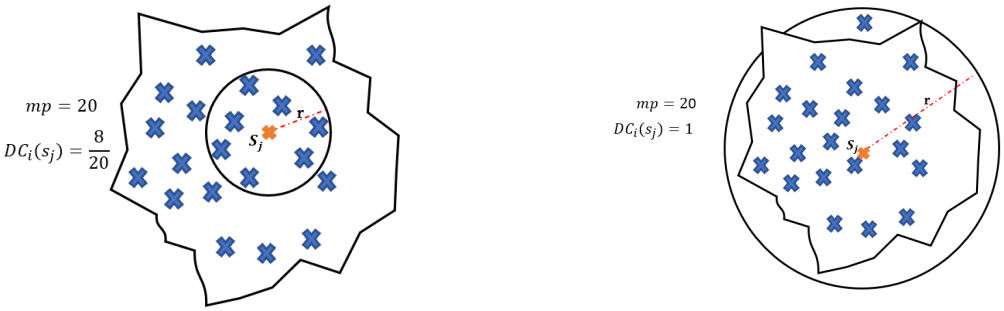


Figure 4. The synthetic point confidence s_j

The number of synthetic points is the number of data in the majority class minus the number of data in the minority class, we noted by N_s :

$$N_s = N_{maj} - N_{min}$$

We assumed the degree of contribution of each synthetic point depends mainly on their degree of confidence. After add this constraint to optimization problem of the DB-FSVM, our approach consists solving the following dual model :

$$\begin{cases} Max \sum_{x_i \in D} \alpha_i - \frac{1}{2} \sum_{x_i \in D} \sum_{x_j \in D} \alpha_i \alpha_j y_i y_j K(x_i x_j) \\ \quad \texttt{Subject to :} \\ \quad \sum_{x_i \in D} \alpha_i y_i = 0 \\ \quad 0 \leq \alpha_i \leq m_i C \quad \forall x_i \in D \ \ \forall i = 1, ..., N \\ \quad 0 \leq \alpha_{s_j} \leq C(s_j) = m_i C \times \frac{Tot(s_j)}{mp} \end{cases} \tag{1}$$

where m_i represents the membershiping degree of the sample x_i to be in its class. Several technic were introduced to calculate m_i [35]. In our case, we use the equation 2 to calculate the parameter m_i. In this sense, suppose there are K classes and lets w_k be the center of the class k obtained by the mean operator and x_i sample

from this class.

$$m_i = \frac{1}{\sum_{j \neq k} (\|x_i - w_k\| / \|x_i - w_j\|)^{\frac{1}{m-1}}} \tag{2}$$

In equation 2, m is a real number greater than 1. We can notice that in the decision function

$$f(x) = \sum_{i=1}^{N_{sv}} \alpha_i^* y_i K(x_i, x) + b, \tag{3}$$

the influence of the α_{s_i} depends on the degree of confidence that it has, that is what it means that the function of the prediction based necessarily on the data of origin.

The mathematical problem that always arises when modifying the constraints of SVM optimization problems is that of solving them with the new constraints. Particularly when the data set is massive, solving classification and regression problems is one of the challenges researchers face in building an efficient maching learning system[36].

Of course, there are quite a few algorithms for solving complex constrained optimization problems, but we will particularly focus on the most commonly used and those that have garnered the most attention recently in the literature, such as Iterative Single Data Algorithms (ISDA) and Sequential Minimal Optimization (SMO) [39], [41], [43], [45]. ISDA algorithms are specifically designed for SVM models, particularly for solving large-scale Support Vector Machine (SVM) problems with the goal of finding the optimal solution efficiently. One of the main features of SVM is the use of kernel functions, which will be described in detail in the next section. A variant of ISDA, the Kernel AdaTron (KA) algorithm, leverages these kernel functions to map data into a high-dimensional space [38], enabling more effective separation of classes in that space [37]. For the second algorithm I've included in the most suitable for solving SVM optimization problems, Platt's SMO algorithm is one of the decomposition algorithms developed in [40], [42], which works on a working set of two data points at a time. Because of the fact that the solution for working set of two can be found analytically, SMO algorithm does not invoke standard QP solvers. Due to its analytical foundation the SMO approach is particularly popular and at the moment the widest used, analyzed and still heavily developing algorithm.

First, up till lately [44], KA appeared to be restricted to classification tasks and second, it "missed" the flower of the robust theoretical framework. KA employs a gradient ascent procedure, and that fact also may have caused some researchers to be suspicious of the challenges posed by gradient ascent techniques in the presence of a perhaps ill-conditioned core array. In [46], for a lacking bias parameter b, the authors derive and demonstrate the equality of two apparently dissimilar ISDAs, namely a KA approach and an unbiased variant of the SMO training scheme [45] when constructing SVMs possessing positive definite kernels. The equivalence is applicable to the classification and regression tasks, and sheds additional insight in these apparently dissimilar methods of learning. Despite the richness of the toolbox set up to solve the quadratic programs from SVM, and with the large amount of data generated by social networks, medical and agricultural fields, etc., the amount of computer memory required for a QP solver from the SVM dual grows hyper-exponentially and additional methods implementing different techniques and strategies are more than necessary.

Kernel Functions of SVM: On the other hand, we have also added a comparison with different versions of the kernel functions in the SVM model, as we are well aware of the crucial role of these functions in improving the prediction rate. Studies on comparisons between SVM kernel functions are numerous and cover various fields. For example, without generalizing, the prediction of the load-bearing capacity of piles [34]. In this study, we have focused on the functions listed in Table 1, as they are the most commonly used and the most efficient.

The choice of parameter values is also a major challenge for these functions. Of course, there are many heuristic methods and metaheuristic techniques. Among these, artificial intelligence methods for local search can solve this type of problem and provide optimal parameter values. In this section, we will explore the most efficient and high-performing intelligent methods for multi-objective optimization problems. These techniques are very successful with complex optimization problems, and all of them are inspired by nature.

Figure 5 shows a diagram of the various metaheuristics that can be used in this kind of problem, including the Bees Algorithm (BeA), Firefly Algorithm (FirA), Particle Swarm Optimization (PSO), Genetic Algorithm

Table 1. Popular Kernel functions of Support Vector Machine

SVM Kernel functions	Mathematic Formula
Linear	$K(x_i, x_j) = x_i.x_j$
Polynomilal	$K(x_i, x_j) = (\gamma \times (x_i.x_j) + C)^d$
Radial Basis Function (RBF)	$K(x_i, x_j) = exp\left(-\gamma \times \|x_i - x_j\|\right)$
Gaussian	$K(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|}{2\sigma^2}\right)$
Sigmoid	$K(x_i, x_j) = \tanh\left(\gamma \times x_i^T.x_j + r\right)$
ANOVA	$K(x_i, x_j) = exp\left(y(x_i - x_j)\right)$

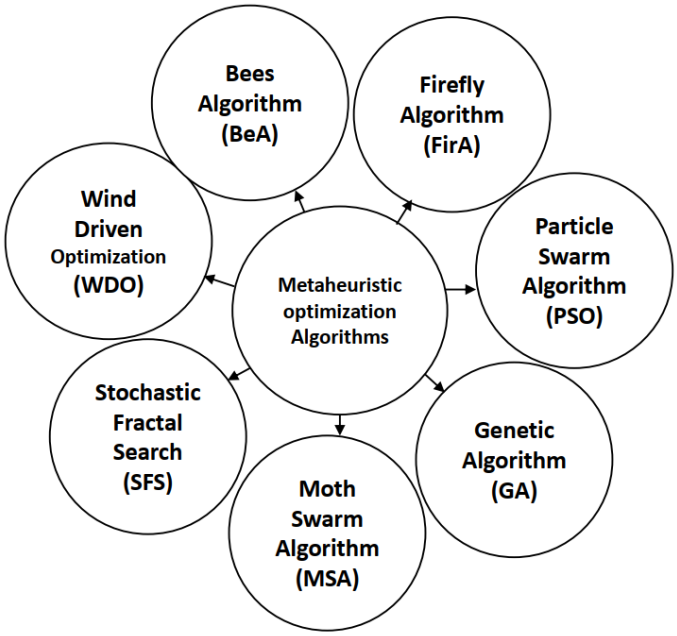


Figure 5. high-performing intelligent methods for multi-objective optimization problems

(GA), Moth Swarm Algorithm (MSA), Stochastic Fractal Search (SFS), and Wind Driven Optimization (WDO) algorithms.

Particle Swarm Optimization (PSO): PSO is a popular metaheuristic optimization method based on swarm intelligence. It draws inspiration from the social behavior of birds flying in groups and fish schools. In this algorithm, the search begins with a population of solutions, each referred to as a ”particle.” Each particle has a velocity and a position in the search space. Throughout the search process, each particle adjusts its position based on its current velocity, its own best-known position found in past iterations, and the best-known position found by the entire swarm [12].

Genetic Algorithm (GA):he genetic algorithm is a well-known evolutionary metaheuristic. It is inspired by biological mechanisms such as Mendel’s laws and the theory of evolution. The algorithm uses vocabulary similar to that in biology and classical genetics, involving terms like genes, chromosomes, individuals, populations, and generations[13].

Stochastic Fractal Search (SFS): Stochastic Fractal Search (SFS) is a global optimization technique based on a population, first proposed by Salimi in 2015 [14]. It belongs to the family of evolutionary algorithms. The technique is based on a fundamental mathematical concept called a ”fractal.” A fractal is a property of an object that exhibits

self-similarity. SFS is a powerful optimization technique leveraging the properties of fractals (self-similarity) and stochastic processes to search large and complex solution spaces efficiently.

the Bees Algorithm (BeA): the Bee Algorithm is a bio-inspired optimization method that mimics the way honeybee colonies search for and exploit food sources. Thanks to a combination of local and global search, it is particularly effective for problems requiring extensive exploration and fine-tuned optimization [15].

Firefly Algorithm (FirA): the Firefly Algorithm is inspired by flashing behavior of firefly insects and immediately attracted the attention of optimization researchers. The Firefly Algorithm is part of a family of swarm intelligence algorithms that have recently shown impressive results in solving optimization problems. The Firefly Algorithm, in particular, is used to solve both continuous and discrete optimization problems[16].

In this paper, we have used the genetic algorithm to solve this problem [35]. In this regards, we have adopted the standard configuration: Initialization(Random uniform) Population size(50), Selection function(Stochastic uniform), Crossover function(Scattered crossover), Crossover probability (0.8), Mutation function(Gaussian mutation), and Max generations(100*Number_Parameters).

Table 2 shows the optimal estimates for each parameter of each kernel function.

Table 2. Estimated values related to the impact of SMOTE-DB-FSVM parameters for each kernel

Parameter Type	Dot	Polynomial	RBF	Neural	Anova
C	1.0e-5	5.0e-5	5.0e-6	5.0e-4	5.0e-4
Convergence Epsilon	0.10	0.01	0.20	0.01	0.01
L_pos	1.30	1.3	1.00	1.30	1.30
L_neg	1.30	1.33	1	1.47	1.47
Kernel Degree	-	3	-	-	3
γ	-	-	2	-	4
Kernel Parameter A	-	-	-	0.01	-
Kernel Parameter B	-	-	-	0.01	-

4. Experimental results

In this section, we assess the performance of the classifier and provide a detailed discussion of the results. Proposed SMOTE-DB-FSVM algorithm was evaluated on two imbalanced datasets to determine its effectiveness. The first dataset, known as the PIMA Indian Diabetes Database [32], provided by the National Institute of Diabetes and Digestive and Kidney Diseases, includes data from 768 patients, with 268 diagnosed as diabetic and 500 as non-diabetic. The second dataset, sourced from a hospital in Frankfurt, Germany [33], contains data from 2000 patients, with 1316 classified as non-diabetic and 684 as diabetic.

Table 3 show the description of two data sets. Table 4 shows the comparison of different classification methods (Naive Bayes, MLP, Knn, AdaBoostM1, SGDClassifier, Nearest Centroid Classifier, and classical SVM) and our SMOTE-DB-FSVM method, evaluated on the PIMA and Germany datasets. We find that the proposed system significantly outperforms all the classification methods considered.

From the Table 4 it is obvious to notice that proposed model SOMTE-DB-FSVM tested on the PIMA dataset gives better results, and he show how our algorithm significantly outperforms classical SVM.

The Table 5 show the comparison of different classifier methods and our SOMTE-DB-FSVM method, evaluated by the Germany data set, also present how much proposed model is capable of distinguishing between classes. From the Table 5 it is obvious to notice that our model SOMTE-DB-FSVM tested on the Germany data set gives better results, and he show how our algorithm significantly outperforms classical SVM.

Another study was conducted to generalize the model across all SVM kernel functions. This study is presented in Table 6, which compares different SVM kernel functions and various oversampling methods on the PIMA dataset. The results demonstrate that the proposed model performs well with all kernel functions and even outperforms the different oversampling methods. Figures 7, 8 and 9 illustrate and validate these results, providing

Table 3. datasets features

	Feature Name	Description	Min val	Max val	Mean
1	Number of pregnancy	Number of times pregnant	0	17	3.85
2	Glucose concentration	2-h oral glucose test (mg/dL)	0	199	120.89
3	Blood Pressure	Diastolic blood pressure (mm Hg)	0	122	69.11
4	Skin thickness	Triceps skin fold thickness (mm)	0	99	20.54
5	Serum Insulin	2-H serum insulin (mu U/mL)	0	846	79.80
6	BMI	Body mass index (kg/m2)	0	67.10	31.99
7	Diabetes Pedigree Function	Diabetes in family history	0.08	2.42	0.47
8	Age	Age in Years	21	81	33.42

Table 4. Comparison between different classification methods on the PIMA dataset

	Performance			
Methode	Accuracy	F1-score	Precision	Recall
Niave Bayes	79.3	67.2	62.50	69.40
MLP	76.6	61.10	57.40	68.30
Knn	80.90	68.40	64.50	66.60
AdaBoostM1	81.10	69.20	70.60	66.50
Dicision Tree	79.90	64.80	72.80	68.70
SGDClassifier	69.03	65.62	66.15	65.83
Nearest Centroid Classifier	66.54	65	66.66	64.89
Classical SVM	79.70	70.7	55.60	62.70
SMOTE-DB-FSVM	91.31	90.33	89.54	90.89

Table 5. Comparison between different classification methods evaluated by the Germany data set

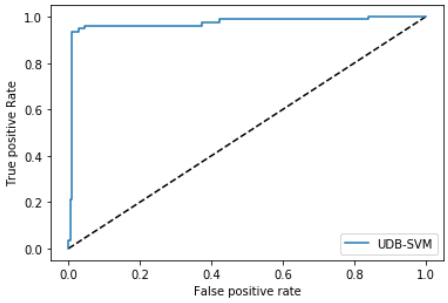
	Performance			
Methode	Accuracy	F1-score	Precision	Recall
Niave Bayes	78.48	75.96	74.17	74.89
MLP	67.73	72.67	51.52	43.64
Knn	81.84	79.85	78.35	79.00
AdaBoostM1	83.02	81.42	79.36	80.22
SGDClassifier	68.40	83.95	52.28	44.81
Nearest Centroid Classifier	73.44	70.75	72.33	71.23
Classical SVM	79.49	78.26	73.52	74.97
SMOTE-DB-FSVM	**96.47**	**95.96**	**96.08**	**96.02**

a visual comparison of the performance of the proposed model against different kernel functions and oversampling methods.

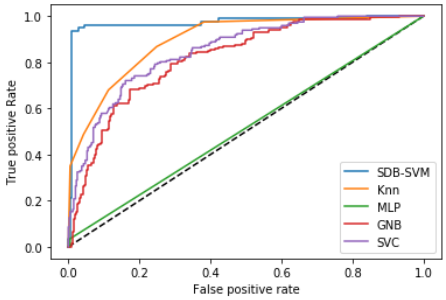
Table 6. Comparisons between different SVM kernel functions and with different oversampling methods on the PIMA dataset

PIMA Dataset					
Classifiers	SVM-linear	SVM-RBF	SVM-Neural	SVM-Polynomial	SVM-ANOV
ADASYN	77.1	78.0	78.0	77.1	77.1
Cluster-Smote	75.9	76.1	75.9	75.9	76.1
MWMOTE	77.4	79.7	67.4	84.3	65.7
SSmote	77.1	79.7	80.5	67.1	63.4
BSmote	79.6	79.7	66.6	84.3	63.5
Random	74.4	73.0	64.4	82.6	63.0
Smote	78.9	73.8	67.2	84.5	63.8
A-SUWO	75.4	74.2	65.4	84.9	64.2
Kmeans Smote	90.1	90.3	90.3	91.1	90.3
Classical SVM	79.7	79.7	79.7	79.7	79.7
SMOTE-DB-FSVM	**91.3**	**92.5**	**91.8**	**91.8**	**91.8**

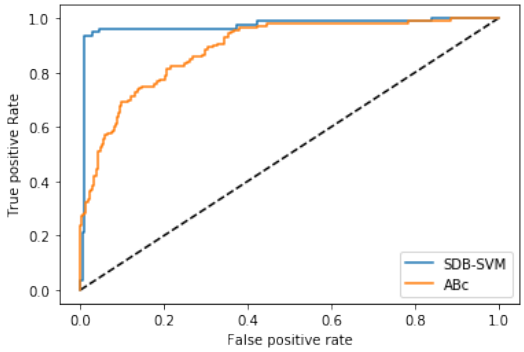
Figures 6a, 6b, and 6c present a comparison of the ROC curves for various classification methods and our SOMTE-DB-FSVM approach, evaluated using the PIMA and Germany diabetes datasets. The ROC curves were employed to compute the AUC values, highlighting the best performance achieved by each classification technique for diabetes prediction. It is evident that the proposed method converges rapidly to optimal results and achieves a higher number of true positives with fewer false positives compared to other classification methods (6c and 6b). This demonstrates the superior performance of SOMTE-DB-FSVM in terms of both accuracy (ACC) and AUC.



(a) ROC Curve of SOMTE-DB-FSVM



(b) ROC Curve SOMTE-DB-FSVM and different classification methods



(c) ROC Curve of SOMTE-DB-FSVM and Abc classifier

Figure 6. ROC Curve of SMOTE-DB-FSVM with different machine learning algorithme

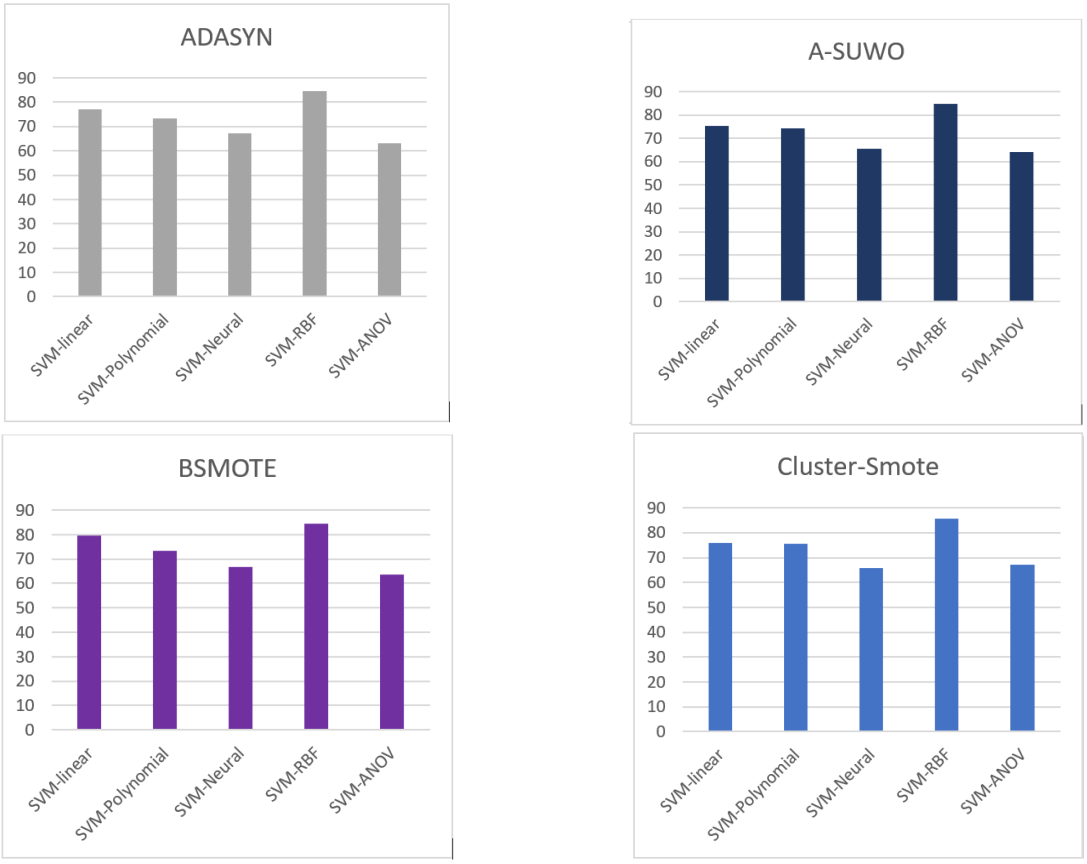


Figure 7. Comparisons between different SVM kernel functions and with different oversampling methods on the PIMA dataset

ROC curve is a performance measurement for classification problem. The 6a, 6b, and 6c present how much a model is capable of distinguishing between classes.

Additionally, it is clear that the SOMTE-DB-FSVM classifier outperforms all other classifiers across all protocols. As seen in Tables 4 and 5, the proposed SOMTE-DB-FSVM model, when tested on the PIMA and Germany diabetes datasets, delivers superior results, with the highest AUC being obtained by our new method. Figures 7, 8, and 9 present a comparison of the performance of various SVM kernel functions applied to the PIMA dataset. The performance of several types of SVM kernels is evaluated to determine which kernel produces the best classification results on this dataset. The figures show the accuracy performance for each SVM kernel function. The comparison highlights the effectiveness of each kernel type in handling the diabetes classification task.

The results show that the FSVM algorithm based on SMOTE's confidence density has clear optimization effects on diabetes prevention as the degree of confidence progressively increases. In addition, the algorithm is clearly capable of handling noisy data as well as unbalanced sets, both of which are challenges for machine learning algorithms. So this hybrid model aims to improve on the DB-FSVM we have already proposed in the literature. The proposed hybrid model shows the quality of the three automatic learning algorithms to avoid the problems that most methods pose. The success of the proposed method based mainly, on the one hand, on the elimination of sound data which has generally reduced the performance of the prediction of conventional models, on the other hand of the balance of classes of all data test.

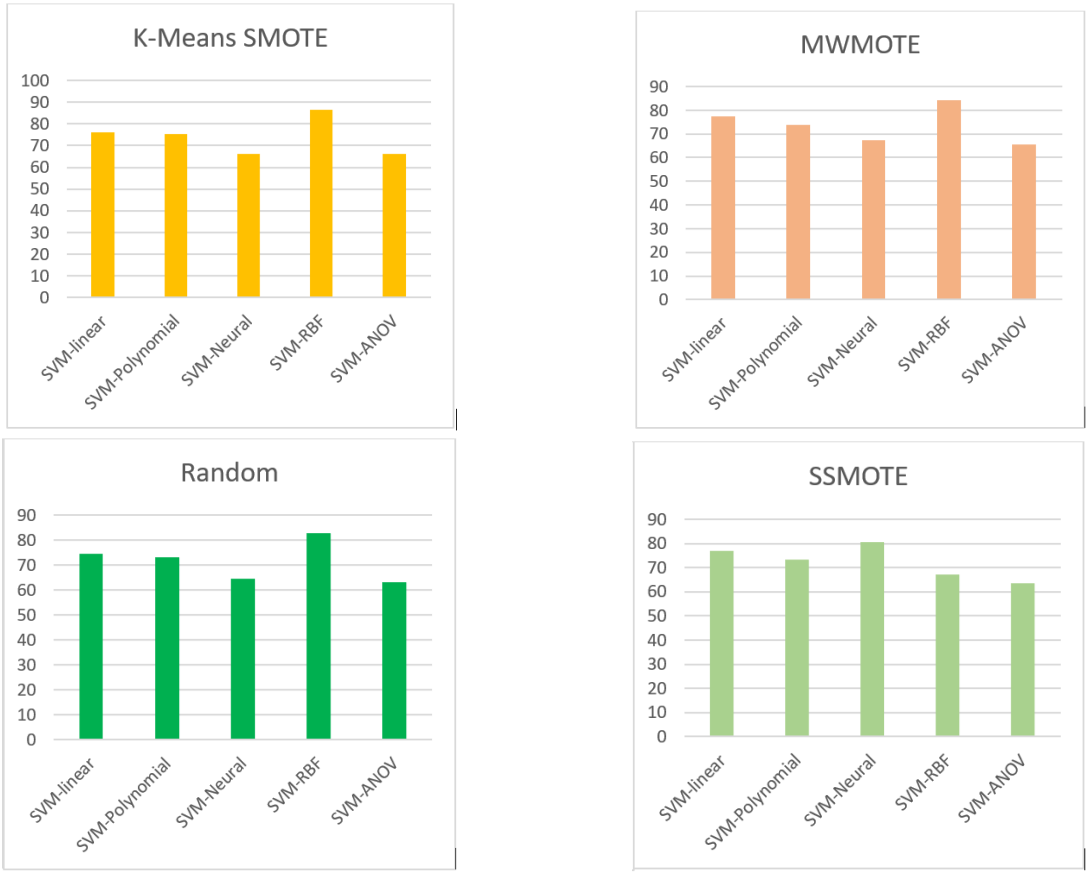


Figure 8. Comparisons between different SVM kernel functions and with different oversampling methods on the PIMA dataset

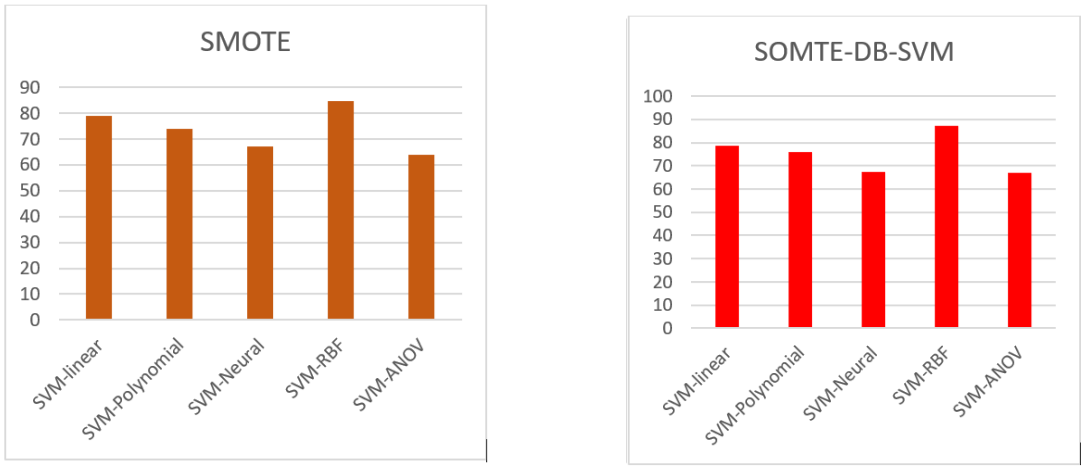


Figure 9. Comparisons between different SVM kernel functions and with different oversampling methods on the PIMA dataset

5. Conclusion

In this work, we introduced a confident SMOTE variant of the density-based fuzzy support vector machine for early diabetes detection. Our proposed method addresses imbalanced datasets through a five main steps: (a) data cleaning, (b) density-based filtering, (c) feature selection to identify the most important attributes, (d) calculation of a confidence score for each point in the minority class, and (e) use of SMOTE to balance the data. We evaluated the proposed system on two imbalanced diabetes datasets, including the PIMA and German datasets, and compared it with well-established classifiers. The experimental results demonstrate the superior performance and efficiency of the new SMOTE-DB-FSVM algorithm. Overall, compared to the conventional SVM model, we observed a significant improvement in performance measures, with the proposed method increasing performance on the dataset by 2.13%, 2.32%, 3.01% and 2.82% for accuracy, F1 score, precision and recall, respectively. Unfortunately, the rigorous adjustment of hyperparameters and the inherent complexity of maintaining the proposed model interpretability further complicate its application, limiting the proposed model generalizability to other diseases, and which would be the objective of future work.

Acknowledgment

This work was supported by Ministry of National Education, Professional Training, Higher Education and Scientific Research (MENFPESRS) and the Digital Development Agency (DDA) and CNRST of Morocco (Nos. Alkhawarizmi/2020/23).

REFERENCES

1. Tang, D., Zhan, Y., & Yang, F. (2024). A review of machine learning for modeling air quality: Overlooked but important issues. Atmospheric Research, 107261.
2. Barbierato, E., & Gatti, A. (2024). The challenges of machine learning: A critical review. Electronics, 13(2), 416.
3. Bian, K., & Priyadarshi, R. (2024). Machine learning optimization techniques: a Survey, classification, challenges, and Future Research Issues. Archives of Computational Methods in Engineering, 1-25.
4. Yadav, A. K., Sagar, D., & Rani, N. (2024, January). A Study on Non-invasive Diabetes Causing Variables and Their Covariance Relationship in Diabetes Prediction Using Machine Learning Algorithms. In International Conference on Smart Computing and Communication (pp. 365-375). Singapore: Springer Nature Singapore.
5. Kumar, B., Negi, H. S., Dimri, S. C., & Rana, D. (2024, March). Prediction of Diabetes Mellitus Disease Using Supervised Machine Learning Models. In 2024 International Conference on Emerging Smart Computing and Informatics (ESCI) (pp. 1-5). IEEE.
6. Shimpi, J. K., Shanmugam, P., & Stonier, A. A. (2024). Analytical model to predict diabetic patients using an optimized hybrid classifier. Soft Computing, 28(3), 1883-1892.
7. Soltanizadeh, S., & Naghibi, S. S. (2024). Hybrid CNN-LSTM for Predicting Diabetes: A Review. Current Diabetes Reviews, 20(7), 77-84.
8. Luo, M., Xiao, F., Chen, Z. Y., Wang, X. K., Hou, W. H., & Wang, J. Q. (2024). A hybrid FSRF model based on regression algorithm for diabetes medical expense prediction. Technological Forecasting and Social Change, 207, 123634.
9. Dohare, S., Pamulaparthy, L., Abdufattokhov, S., Naga Ramesh, J. V., El-Ebiary, Y. A. B., & Thenmozhi, E. (2024). Enhancing Diabetes Management: A Hybrid Adaptive Machine Learning Approach for Intelligent Patient Monitoring in e-Health Systems. International Journal of Advanced Computer Science & Applications, 15(1).
10. Douzas, G., Bacao, F., & Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. Information sciences, 465, 1-20.
11. Liang, X. W., Jiang, A. P., Li, T., Xue, Y. Y., & Wang, G. T. (2020). LR-SMOTE—An improved unbalanced data set oversampling based on K-means and SVM. Knowledge-Based Systems, 196, 105845.
12. El Moutaouakil, K., Cheggour, M., Chellak, S., & Baïzri, H. (2021, July). Metaheuristics Optimization Algorithm to an Optimal Moroccan Diet. In 2021 7th Annual International Conference on Network and Information Systems for Computers (ICNISC) (pp. 364-368). IEEE.
13. Ahourag, A., El Moutaouakil, K., Chellak, S., Baizri, H., & Cheggour, M. (2022, May). Multi-criteria optimization for optimal nutrition of Moroccan diabetics:* Note: Sub-titles are not captured in Xplore and should not be used. In 2022 International Conference on Intelligent Systems and Computer Vision (ISCV) (pp. 1-6). IEEE.
14. Abdellatif, E.O., Karim, E.M., Hicham, B. et al. Intelligent Local Search for an Optimal Control of Diabetic Population Dynamics. Math Models Comput Simul 14, 1051–1071 (2022). https://doi.org/10.1134/S2070048222060047.
15. El Moutaouakil, K., El Ouissari, A., Hicham, B., Saliha, C., & Cheggour, M. (2022). Multi-objectives optimization and convolution fuzzy C-means: control of diabetic population dynamic. RAIRO-Operations Research, 56(5), 3245-3256.

16. El Moutaouakil, K., & Touhafi, A. (2020, November). A New Recurrent Neural Network Fuzzy Mean Square Clustering Method. In 2020 5th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech) (pp. 1-5). IEEE.
17. El Moutaouakil, K., el Ouissari, A., Touhafi, A., & Aharrane, N. (2020, November). An Improved Density Based Support Vector Machine (DBSVM). In 2020 5th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech) (pp. 1-7). IEEE.
18. Kumari, V. Anuja, and R. Chitra. "Classification of diabetes disease using support vector machine." International Journal of Engineering Research and Applications 3.2 (2013): 1797-1801.
19. Hassan, M. M., and Amiri, N. Classification of Imbalanced Data of Diabetes Disease Using Machine Learning Algorithms. International Conference on Theoretical and Applied Computer Science and Engineering (ICTACSE, 2019) (2019), 21(81), 33-24.
20. N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357, ISSN 10769757, doi: 10.1613/ jair.953 .
21. H. Han, W.-Y. Wang, B.-H. Mao, Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning, Advances in intelligent computing 17 (12) (2005) 878–887, ISSN 1941-0506, doi: 10.1007/11538059 91 .
22. D. A. Cieslak, N. V. Chawla, A. Striegel, Combating imbalance in network intrusion datasets, in: IEEE International Conference on Granular Computing, 2006, IEEE, ISBN 1-4244-0134-8, 732–737, doi: 10.1109/GRC.2006.1635905 , 2006.
23.] I. Nekooeimehr, S. K. Lai-Yuen, Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets, Expert Systems with Applications 46 (2016) 405–416, ISSN 09574174, doi: 10.1016/j.eswa.2015.10.031 .
24. W.-C. Lin, C.-F. Tsai, Y.-H. Hu, J.-S. Jhang, Clustering-based undersampling in class-imbalanced data, Information Sciences 409-410 (2017) 17–26, ISSN 0020-0255, doi: 10.1016/j.ins.2017.05.008 .
25. Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. ICT Express.
26. Tigga, Neha Prerna, and Shruti Garg. "Prediction of type 2 diabetes using machine learning classification methods." Procedia Computer Science 167 (2020): 706-716.
27. Shuja, M., Mittal, S., and Zaman, M. (2020). Effective prediction of type ii diabetes mellitus using data mining classifiers and SMOTE. In Advances in computing and intelligent systems (pp. 195-211). Springer, Singapore.
28. Devi, R. D. H., Bai, A., and Nagarajan, N. (2020). A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms. Obesity Medicine, 17, 100152.
29. Uddin, M. A., Islam, M. M., Talukder, M. A., Hossain, M. A. A., Akhter, A., Aryal, S., & Muntaha, M. (2024). Machine learning based diabetes detection model for false negative reduction. Biomedical Materials & Devices, 2(1), 427-443.
30. Mohammed, R. (2024). FCM-CSMOTE: Fuzzy C-Means Center-SMOTE. Expert Systems with Applications, 248, 123406.
31. A. El Ouissari and K. El Moutaouakil, "Density based fuzzy support vector machine: application to diabetes dataset," Math. Model. Comput. 8 (4), 747–760 (2021). https://doi.org/10.23939/mmc2021.04.747
32. https://www.kaggle.com/uciml/pima-indians-diabetesdatabase
33. https://www.kaggle.com/johndasilva/diabetes
34. Wang, Q., Moreno-Martínez, Á., Muñoz-Marí, J., Campos-Taberner, M., & Camps-Valls, G. (2023). Estimation of vegetation traits with kernel NDVI. ISPRS Journal of Photogrammetry and Remote Sensing, 195, 408-417.
35. El Moutaouakil, K., & El Ouissari, A. (2023). Opt-RNN-DBFSVM: Optimal recurrent neural network density based fuzzy support vector machine. RAIRO-Operations Research, 57(5), 2493-2517.
36. K. El Moutaouakil, M. Roudani, & A. El Ouissari, Optimal Entropy Genetic Fuzzy-C-Means SMOTE (OEGFCM-SMOTE). *Know. Bas. Sys.* 262 (2023) 110235.
37. J. K. Anlauf, M. Biehl, The AdaTron – an adaptive perceptron algorithm. *Europhysics Letters* 10(7) (1989) 687–692.
38. T.-T. Frieß, Cristianini, N., Campbell, I. C. G., The Kernel-Adatron: a Fast and Simple Learning Procedure for Support Vector Machines. *In Shavlik, J., editor, Proceedings of the 15th International Conference on Machine Learning, Morgan Kaufmann, San Francisco, CA* (1998) 188–196.
39. T.-M. Huang, V. Kecman, Bias Term b in SVMs Again, *Proc. of ESANN 2004, 12th European Symposium on Artificial Neural Networks Bruges, Belgium*, 2004.
40. T. Joachims, Making large-scale svm learning practical. advances in kernel methods-support vector learning. *http://svmlight. joachims.org/* (1999).
41. V. Kecman, M. Vogt, T.-M. Huang, On the Equality of Kernel AdaTron and Sequential Minimal Optimization in Classification and Regression Tasks and Alike Algorithms for Kernel Machines, *Proc. of the 11 th European Symposium on Artificial Neural Networks, ESANN, , Bruges, Belgium* (2003) 215–222.
42. E. Osuna, R. Freund, F. Girosi, An Improved Training Algorithm for Support Vector Machines. *In Neural Networks for Signal Processing VII, Proceedings of the 1997 Signal Processing Society Workshop*, (1997) 276–285.
43. J. Platt, Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, *Microsoft Research Technical Report MSR-TR-98-14,* (1998).
44. Veropoulos, K., Machine Learning Approaches to Medical Decision Making, *PhD Thesis, The University of Bristol, Bristol, UK* (2001).
45. M. Vogt, SMO Algorithms for Support Vector Machines without Bias, Institute Report, *Institute of Automatic Control, TU Darmstadt, Darmstadt, Germany, (Available at http://www.iat.tu-darmstadt.de/vogt)* (2002).
46. V. Kecman, T. M.Huang, & M. Vogt, Iterative single data algorithm for training kernel machines from huge data sets: Theory and performance. *Support vector machines: Theory and Applications,* (2005) 255-274.