Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 20 No. 8 (2024) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v20i08.47765

PAPER

# Diabetes Prediction: Optimization of Machine Learning through Feature Selection and Dimensionality Reduction

**Abd Allah Aouragh[1](✉), Mohamed Bahaj[1], Fouad Toufik[2]**

[1]MIET Laboratory, Faculty of Sciences and Techniques, Hassan 1st University, Settat, Morocco

[2]Computer Sciences Laboratory, Higher School of Technology, Mohammed V University, Sale, Morocco

a.aouragh@uhp.ac.ma

**ABSTRACT**

Diabetes, a pervasive global health concern, presents diagnostic challenges due to its nuanced onset and far-reaching implications. Traditional diagnostic approaches, reliant on time-consuming assessments, necessitate a paradigm shift towards more efficient methodologies. In response, this study introduces a diagnostic support system leveraging the power of optimized machine learning algorithms. Addressing class imbalance within a dataset comprising 768 records, our methodology intricately weaves together feature selection, dimensionality reduction techniques, and grid search optimization. Specifically, the Extra Trees model, fine-tuned via grid search, emerges as the most potent, showcasing remarkable performance metrics: an accuracy score of 92.5%, an F1-score of 93.7%, and an AUC-ROC of 92.47%. These findings underscore the pivotal role of machine learning in reshaping diabetes diagnosis, offering transformative possibilities for global healthcare enhancement.

**KEYWORDS**

diabetes, machine learning, balancing, feature selection, dimensionality reduction, grid search

## 1 INTRODUCTION

Diabetes is a persistent health disorder that manifests when there is deficient insulin secretion from the pancreas or when the body faces challenges in utilizing the insulin it generates efficiently. Insulin plays a crucial role in managing blood sugar levels. Uncontrolled diabetes often leads to hyperglycemia, which is marked by high blood glucose levels [1, 2].

According to the latest statistics released by the World Health Organization (WHO), diabetes has become a serious metabolic challenge, spreading on a global scale and generating growing public awareness [2]. The proportion of people affected by this disease has climbed dramatically, from 108 million in 1980 to an alarming 422 million in 2014 [2].

The nuanced manifestations of diabetes, including heightened thirst, recurrent urination, enduring fatigue, and unexplained weight loss, are indicative of

Aouragh, A.A., Bahaj, M., Toufik, F. (2024). Diabetes Prediction: Optimization of Machine Learning through Feature Selection and Dimensionality Reduction. *International Journal of Online and Biomedical Engineering (iJOE)*, 20(8), pp. 100–114. https://doi.org/10.3991/ijoe.v20i08.47765

Article submitted 2024-01-05. Revision uploaded 2024-02-22. Final acceptance 2024-02-23.

underlying disruptions in carbohydrate metabolism [1, 2]. Despite being frequently overlooked, these initial warning signs serve as harbingers of potential severe complications. In certain instances, they may escalate into cardiovascular diseases, renal impairments, ocular complications and in more extreme cases, necessitate lower-limb amputations [2]. It is crucial to recognize and address these early indicators promptly, as they offer valuable insights into the potential trajectory of the disease and provide opportunities for preventive interventions [1, 2]. In addition to the consequences for individual health, diabetes also causes considerable financial pressure on the world's healthcare systems because of its high treatment and management costs [3]. In this context, the necessity of an early diagnosis tool becomes imperative to mitigate these consequences. However, the diagnosis of diabetes presents inherent challenges, requiring extensive laboratory analyses and rigorous clinical assessments encompassing blood glucose levels, insulin functionality, and overall metabolic health [4]. Given the intricate nature of the disease and the potential individual variations in response, a nuanced and multifaceted diagnostic approach becomes imperative [4]. This complexity underscores the importance for researchers and healthcare professionals to employ new diagnostic methodologies, ensuring a comprehensive and precise evaluation for the effective identification and management of diabetes [4, 5].

Machine learning (ML), a significant technological leap, has become a driving force in the field of medicine. This formidable branch of artificial intelligence (AI), harnessing intricate algorithms to scrutinize vast datasets, has unveiled unprecedented avenues for understanding and addressing diseases [6]. When applied to the domain of diabetes, machine learning becomes a catalyst with the capability to enhance early detection accuracy and optimize healthcare management. By leveraging complex algorithms to analyze extensive datasets, machine learning not only facilitates more nuanced insights into the dynamics of diabetes but also paves the way for more tailored and effective healthcare strategies [6, 7]. In our study, we explored this convergence between artificial intelligence and medicine using machine learning algorithms on an unbalanced dataset of diabetic patients. This innovative approach, in conjunction with other techniques such as data balancing, feature selection, dimensionality reduction, and hyperparameter optimization, has produced very promising results. The achievements of our research demonstrate the revolutionary power of machine learning to improve diagnostics and enhance healthcare on a large scale. The primary contributions of this study can be encapsulated in the points listed below:

- Integrating machine learning into diabetes diagnosis
- Handling unbalanced dataset
- Optimizing feature selection and dimensionality reduction
- Optimizing hyperparameters to achieve high-performance
- Compared with the state-of-the-art, the suggested system attains exceptional performance levels

## 2 RELATED WORK

In recent years, there has been a constant evolution and exponential growth in medical research employing machine learning. Researchers harness machine learning tools to process vast and intricate datasets from diverse sources, such as electronic medical records, laboratory test results, and clinical reports. This extensive data fuels the training of machine learning algorithms, facilitating the discovery of concealed patterns and relationships [8, 9]. This innovative approach holds the promise for scientists to construct forecasting models of diabetes, devise tailored

treatments, and advance the frontiers of medical research [10]. In this context, diabetes has emerged as a focal point, contributing a crucial dimension to the dynamic intersection between machine learning and medical insight.

In this context, Chaki et al. [11] carried out a comprehensive systematic review, analyzing 107 publications. Their study underscores the influence of progress in machine learning and artificial intelligence, showcasing superior performance in early identification and automated diagnosis of diabetes compared to conventional manual methods. The review offers a detailed examination of techniques for diabetes detection, diagnosis, and self-management, encompassing approaches to preprocessing data, extracting features, conducting machine learning determination, and assessing performance metrics. Alanazi et al. [12] conducted a thorough literature review to offer a detailed analysis of artificial intelligence and machine learning techniques employed in diabetes management. Their analysis delves into the advantages and limitations of utilizing these techniques in diabetes management, identifying areas that necessitate future research. The review illustrates the groundbreaking capacity of AI and ML in revolutionizing diabetes management, facilitating more precise and efficient diagnosis and treatment. It also underscores the importance of addressing challenges like data quality, system transparency, and ethical considerations through further research.

Al-Zebari et al. [13] carried out an exhaustive comparative analysis of machine learning techniques for diabetes detection. The study investigates various methods, including techniques like logistic regression, decision trees, support vector machines, discriminant analysis, k-nearest neighbors, and ensemble methods, employing the MATLAB classification learner tool. A total of 24 classifiers were assessed through 10-fold cross-validation, registering an average classification accuracy varying from 65.5% to 77.9%. Logistic regression yielded the highest accuracy at 77.9%; on the other hand, the coarse Gaussian SVM technique exhibited the lowest performance at 65.5%. Sonar et al. [14] devised a sophisticated machine learning-based system focused on data processing to predict the onset of diabetes in patients, enabling timely intervention. They constructed classification models, including artificial neural networks, decision trees, support vector machines, and naïve Bayes. Their findings demonstrate substantial accuracy, with the decision tree achieving 85%, naïve Bayes at 77%, and SVM at 77.3%, underscoring the efficacy of their approach in forecasting diabetes risk. Kopitar et al. [15] investigated contemporary methods for detecting type 2 diabetes, predominantly relying on multivariate regression. Their analysis compares the efficiency of machine learning prediction techniques (LightGBM, Glmnet, XGBoost, and RF) with that of conventional regression models commonly applied in forecasting undetected diabetes cases. In terms of mean RMSE, the basic regression model outperformed others, with the lowest value of 0.838. RF, LightGBM, Glmnet, and XGBoost achieved the subsequent values: 0.842, 0.846, 0.859, and 0.881, respectively. García-Ordás et al. [16] addressed the growing challenge of diabetes as a chronic disease and the crucial need for early diagnosis. Their study focused on a pipeline that employs deep learning techniques for predicting diabetes in individuals. Applying data augmentation through a variational autoencoder (VAE), enhancing features with a sparse autoencoder (SAE), and conducting classification using a convolutional neural network (CNN), the approach demonstrated an impressive accuracy rate of 92.31% on the Pima Indians Diabetes database, signifying a noteworthy improvement over existing methods. Khanam et al. [17] studied diabetes prognosis, leveraging the Pima Indian diabetes dataset. Their research, employing various methodologies, including data analysis and machine learning techniques, found that a hybrid model incorporating support vector machines and logistic regression outperformed others. Neural networks with a structure of two hidden layers achieved a notable 88.6% accuracy, highlighting the critical role of machine learning advancements in enhancing

diagnostic techniques for early diabetes detection. Ashraf Uddin et al. [18] developed a machine learning model incorporating decision trees, k-nearest neighbors, logistic regression, random forest, naïve Bayes, and support vector machine methods. Utilizing data preprocessing and the SMOTE technique for dataset balancing, the random forest algorithm achieved notable accuracy of 97% and 80% on the 2019 and Pima Indian datasets, respectively, emphasizing the significance of balanced datasets in minimizing false negatives. Febrian et al. [19] conducted an in-depth investigation using machine learning methods to analyze and evaluate the efficiency of the naïve Bayes and k-nearest neighbors techniques in predicting diabetes. Utilizing the Pima Indians Diabetes Database dataset, their results unequivocally demonstrated the superiority of the naïve Bayes model, achieving a noteworthy accuracy of 78.57%. Sihlangu et al. [20] investigated diverse machine learning approaches, such as logistic regression, stochastic gradient descent, CN2 rule, and support vector machines, employing the Orange data science tool. Their primary aim was to forecast diabetes using the PIMA Indian Diabetes dataset. The CN2 rule induction approach was demonstrated to be the most efficient, attaining an accuracy of 80.7%. Mousa et al. [21] assessed three models: Convolutional Neural Network (CNN), Random Forest (RF), and Long Short-Term Memory (LSTM) for diabetes recognition using the Pima Indian database. The LSTM demonstrated superior performance with a maximum accuracy of 85%, while the RF and CNN, though promising, exhibited slightly lower performance. Building on prior research, it is clear that machine learning represents an innovative approach in the realm of medical predictive modeling, specifically for the identification and diagnosis of diabetes. Researchers have investigated a diverse array of machine learning techniques to anticipate this complex disease, recognizing that no single universal method is applicable to all cases. In our study, five machine learning algorithms were utilized for diabetes prediction.

## 3 MATERIALS AND METHODS

### 3.1 Dataset

In the realm of machine learning, datasets play a critical role, serving as the foundation for algorithm learning and performance improvement. Our study opted for the well-established Pima Indians Diabetes Database [22], sourced from the Pima Native American community in the southwestern United States. This dataset encompasses vital diabetes-related information, including attributes like glucose concentration, tricipital skinfold thickness, blood pressure, and body mass index (BMI), among others. Comprising 768 entries, each corresponding to a patient record, this dataset serves as a fundamental resource for evaluating diabetes indicators. The detailed structure of the dataset is delineated in Table 1.

**Table 1.** Dataset features

| Num | Attribute | Description |
|---|---|---|
| 1 | Pregnancies | Count of pregnancies |
| 2 | Glucose | Plasma glucose levels |
| 3 | BloodPressure | Diastolic blood pressure (mm Hg) |
| 4 | SkinThickness | Triceps skin fold thickness (mm) |

*(Continued)*

**Table 1.** Dataset features *(Continued)*

| Num | Attribute | Description |
|---|---|---|
| 5 | Insulin | 2-Hour serum insulin (mu U/ml) |
| 6 | BMI | Body mass index |
| 7 | DiabetesPedigreeFunction | Diabetes pedigree function |
| 8 | Age | Age (years) |
| 9 | Outcome | Target: 1 = diabetic, 0 = non diabetic |

The dataset comprised 65.1% (500 cases) of non-diabetic and 34.9% (268 cases) of diabetic instances. Imbalances in machine learning models may lead to suboptimal predictions, reduced minority class recall, and overfitting for the majority class. To overcome these challenges, employing dataset balancing techniques is essential for the accurate evaluation of diabetes prediction models.

## 3.2 Data preprocessing

Data preprocessing is the first critical phase in the development of machine learning algorithms, and it is extremely vital for guaranteeing the quality, reliability, and performance of prediction algorithms [23]. For our Pima Indians Diabetes Database dataset, we performed the following techniques during this preliminary phase:

- Outlier Handling: To bolster the robustness of our model, we applied the "Replace with Thresholds IQR" method for handling outliers. This approach involves replacing extreme values with thresholds based on the Interquartile Range (IQR), contributing to a more resilient and reliable model.
- Normalization: Ensuring that certain features do not disproportionately influence the model due to differing scales, we employed the RobustScaler for normalization. This technique enhances the convergence of the model by standardizing feature scales.
- Handling Imbalanced Data: Prevent bias and ensure that the model is not influenced by the prevalence of a particular class. As part of our study, we rigorously evaluated different class imbalance handling strategies, including resampling (RESAMPLE), SMOTE (Synthetic Minority Over-sampling Technique), and ADASYN (Adaptive Synthetic Sampling) [24]. After thorough analysis, it was found that the RESAMPLE method outperformed the others, providing the best results. Consequently, this approach was chosen to effectively overcome the class imbalance present in the dataset.

## 3.3 Feature selection

Feature selection is a crucial phase in machine learning projects, involving the careful selection of informative attributes to improve model efficiency. This step aims to reduce data dimensionality while retaining feature relevance for better model generalization [25, 26]. In our study, we utilized two techniques, k-best and variance threshold, chosen for their adaptability and effectiveness in enhancing machine learning models.

- KBest, or k-best feature method, utilizes a univariate approach to evaluate each feature independently. It calculates a correlation measure between each attribute and the target variable, often employing the $\chi 2$ (chi-square) test for categorical

target variables. This statistical measure assesses the independence of variable distributions, with a high score indicating strong feature-dependence for prediction. The KBest method with the $\chi 2$ test excels in identifying informative features [25, 26]. In our study, it selected the five most informative features.

– The variance threshold technique assesses each feature's variance, removing those below a preset threshold. Features with low variance, indicating little variation among samples, are deemed less informative for prediction. This method effectively eliminates constant or quasi-constant features, simplifying models and reducing overfitting risk. By enhancing computational efficiency through feature reduction, it improves machine learning model effectiveness [25, 26]. In our study, a threshold of 0.40 was applied for feature selection using the variance threshold method.

## 3.4 Dimensionality reduction

– Dimensionality reduction simplifies complex models by decreasing dataset variables while retaining essential information and trends. This is vital for datasets with many dimensions to prevent computational inefficiencies, overfitting, and visualization challenges. In our study, we applied t-SNE (t-distributed Stochastic Neighbor Embedding) and MDS (Multidimensional Scaling) techniques to uncover underlying structures in the Pima Indians Diabetes Database and aid result interpretation [25, 27].
– t-SNE (t-distributed Stochastic Neighbor Embedding) is a dimensionality reduction technique known for its effectiveness in representing complex, high-dimensional data in a reduced space. It focuses on preserving local similarities and non-linear relationships between samples by creating joint probability distributions and adjusting them to minimize divergence [25, 27]. In our study, applying t-SNE with five dimensions aims to reveal intricate relationships in the Pima Indians Diabetes Database, offering a meaningful representation of underlying structures.
– MDS (Multidimensional Scaling) is a technique used for dimensionality reduction, aiming to visually represent the structure of similarities or dissimilarities between observations. It works by creating an initial distance matrix based on similarity measures like correlations or Euclidean distances, then assigning positions in a reduced-dimensional space to each observation to best reflect the initial distances. MDS excels at capturing complex structures in the data, offering an intuitive visualization [27]. In our study, MDS with five dimensions was chosen to explore the underlying structure of the data following various preprocessing steps.

## 3.5 Machine learning algorithms

In our study, we employed five well-established machine learning (ML) algorithms: K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest, Extra Trees, and Gradient Boosting. These algorithms were chosen based on extensive research indicating their strong performance in datasets similar to ours [27, 28]. They have demonstrated effectiveness in diagnosing health disorders and excel in handling classification tasks, especially in scenarios with complex relationships between variables. The widespread recognition of these algorithms in the machine learning community underscores their suitability for our study [27, 28].

– K-Nearest Neighbors (KNN) is a simple and adaptable supervised learning algorithm that evaluates data point proximity in feature space. It classifies new observations based on their k-nearest neighbors in the training set, usually using

Euclidean distance. Despite its simplicity and ability to handle complex data structures without distribution assumptions, KNN may be sensitive to noise or high-dimensional feature spaces, which can affect its performance [27, 28].

- Support Vector Machines (SVM) is a powerful supervised learning method for complex classification tasks. SVM seeks the best hyperplane to separate classes in feature space, maximizing the margin between the nearest points and the hyperplane. It excels with high-dimensional data and can handle non-linear datasets using kernel functions, enabling classification in complex feature spaces [27, 28].
- Random Forest is an ensemble learning method that constructs multiple decision trees during training. Each tree is trained on a random subset of the data, and their predictions are aggregated to make a final forecast. This technique utilizes tree diversity to mitigate overfitting and enhance model generalization. Each tree contributes to the decision process by voting for a class, with the majority class determining the final prediction [27, 28].
- Extra Trees is a variant of Random Forest that utilizes a group of decision trees, similar to Random Forest but with a different construction process. Unlike Random Forest, which selects the best threshold for node division from a random subset of features, Extra Trees employs entirely random thresholds for each feature. This approach aims to maximize ensemble tree diversity by introducing more randomness into the tree-building process, enhancing the model's ability to generalize on unseen data [25, 29].
- Gradient Boosting is an ensemble learning method that improves predictive model efficiency by integrating multiple weaker models' predictions. Unlike Random Forest, it focuses on correcting model errors iteratively, refining accuracy by reducing residual errors [27, 28].

### 3.6 Grid search and K-fold cross validation

Hyperparameter optimization and cross-validation are crucial techniques for improving machine learning model performance. In our study, we utilized Grid Search and k-fold cross-validation to achieve more accurate and generalizable models. Grid Search systematically explores different combinations of predefined hyperparameters to identify the optimal set, particularly beneficial when model performance depends heavily on specific hyperparameter values [25, 30]. Additionally, k-fold cross-validation is a robust evaluation strategy that divides the dataset into k-folds, using k-1 folds for training and one-fold for validation in each iteration. This process provides a reliable assessment of model performance across the entire dataset, reducing the risk of overfitting or underfitting specific dataset characteristics [25, 30]. We chose to implement k-fold cross-validation with k = 10 in our study. This choice balanced the need for a robust assessment of model performance with computational resource limitations. Moreover, k = 10 is commonly considered a standard choice in the machine learning research community, offering a good balance between accuracy and computational efficiency.

## 4 METHODOLOGY AND EVALUATION METRICS OVERVIEW

### 4.1 Global overview

In this study on diabetes prediction using machine learning, we followed a systematic methodology, starting with the retrieval of the Pima Indians Diabetes Database dataset. After initial data preprocessing, including tasks like splitting and normalization, we addressed the data imbalance using the Resample technique. To identify significant

attributes, we compared KBest (chi2 test) and variance threshold for feature selection and t-SNE and MDS for dimensionality reduction. Diverse machine learning techniques were evaluated, with hyperparameter optimization through grid search. Our methodology also employed k-fold cross-validation (k = 10) for robust model evaluation, reducing the risk of overfitting. Figure 1 visually outlines our study's design approach.

For our project's hardware and software setup, we selected a computer featuring an AMD Ryzen 7 5700G processor and a Radeon graphics card, providing robust computing capabilities for machine learning assignments. Jupyter was our chosen development environment, allowing us to generate interactive notebooks for visualizing model outcomes. We utilized the Python programming language along with the Matplotlib, Pandas, and Scikit-Learn libraries due to their versatility, user-friendly nature, and extensive ecosystem dedicated to machine learning.

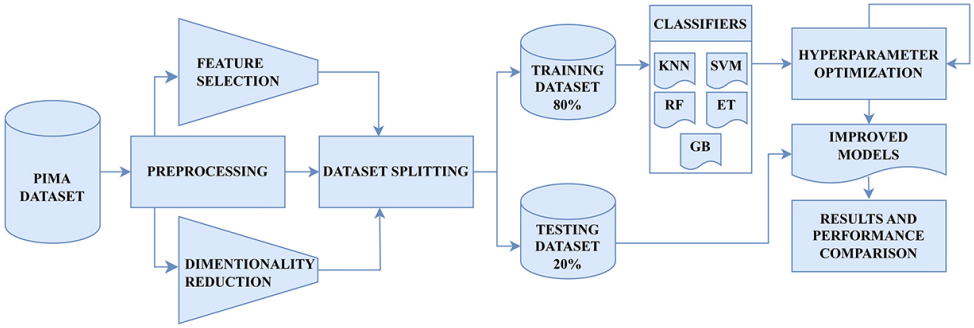


**Fig. 1.** Suggested methodology

### 4.2 Evaluation metrics

Evaluating the efficacy of machine learning algorithms is usually carried out through the confusion matrix, an essential tool for understanding the quality of model predictions [31]. This matrix divides the results into four main components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). From the confusion matrix generated by our machine learning models, we have calculated a set of metrics crucial to assessing their performance. These metrics include:

- Accuracy: the fraction of correct forecasts out of the total predictions [31].
- Precision: the fraction of true positive forecasts among the total positive predictions [31].
- Recall: measures the proportion of true positive instances identified by the model relative to the total number of actual positive instances [31].
- F1-score: integrates precision and recall into one single value. It provides a balance between these two measures [31].
- AUC-ROC: Area Under the Receiver Operating Characteristic Curve, measures the area under the curve when plotting the True Positive Rate against the False Positive Rate [31].

## 5 RESULTS AND DISCUSSION

This part describes and discusses the findings achieved in our study of diabetes prediction utilizing the aforementioned machine learning algorithms. The overall

efficiency of each algorithm, including the optimization of hyperparameters using grid search and cross-validation, is detailed in Tables 2–6 in terms of accuracy, recall, precision, F1-score, and area under the ROC curve (AUC-ROC). Abbreviations used in the tables include KNN (K-Nearest Neighbors), SVM (Support Vector Machines), RF (Random Forest), ET (Extra Trees), GB (Gradient Boosting), and GS (Grid Search).

The diverse findings reveal the substantial influence of the approaches we advocate on improving the performance of all classification algorithms. For the initial dataset, whose results are presented in Table 2, Extra Trees recorded the best accuracy with a value of 78.36%. Regarding precision, the best value was recorded by SVM with a score of 84.62%. On the other hand, Gradient Boosting dominated the following best values: a recall of 71.59%, an F1-score of 71.19%, and an AUC-ROC of 76.70%. After data balancing and t-SNE application, Table 3 reveals that the optimized SVM {'C': 1, 'gamma': 1, 'kernel': 'rbf'} dominated the following best metrics: accuracy of 90.50%, precision of 93.62%, F1-score of 90.26%, and AUC-ROC of 90.53%. The best recall was recorded by the optimized KNN {'algorithm': 'auto', 'metric': 'manhattan', 'n_neighbors': 7, 'weights': 'distance'} with a value of 95.05%. As a result of applying the Multidimensional Scaling (MDS) after balancing, Table 4 reveals that the optimized Extra Trees {'n_estimators': 100, 'min_samples_split': 4, 'criterion': 'gini', 'min_samples_leaf': 1} stands out with the following best scores: an accuracy of 92.00%, a precision of 89.72%, a recall of 95.05%, an F1-score of 92.31%, and an AUC-ROC of 91.97%. After applying data balancing in conjunction with the KBest method, Table 5 reveals the predominance of optimized Extra Trees {'n_estimators': 200, 'min_samples_split': 4, 'criterion': 'gini', 'min_samples_leaf': 1}, demonstrating exceptional performance, including accuracy of 92.50%, precision of 90.57%, recall of 97.05%, F1-score of 93.70%, and AUC-ROC of 92.47%. Furthermore, the combination of data balancing with variance threshold, as illustrated in Table 6, highlights the superior performance of the optimized Random Forest {'n_estimators': 500, 'min_samples_split': 2, 'criterion': 'entropy', 'min_samples_leaf': 1}. Results achieved include an accuracy of 91.50%, a precision of 87.61%, a recall of 97.04%, an F1-score of 92.08%, and an AUC-ROC of 90.64%.

These results underscore a significant enhancement across all assessed metrics, illustrating the positive influence of various optimization phases on algorithmic performance, resulting in an improvement of up to 14.27%. Notably, the top-performing model emerged from utilizing the balanced dataset with the KBEST approach. Through meticulous optimization, conducted via grid search and cross-validation, Extra Trees achieved remarkable results, boasting an accuracy of 92.50% and an F1-score of 93.70%. These findings surpass the performance of other techniques, including the balanced dataset with t-SNE, which achieved an accuracy of 90.50% and an F1-score of 90.26%, MDS with an accuracy of 92.00% and an F1-score of 92.31%, and variance threshold with an accuracy of 91.50% and an F1-score of 92.08%. Figure 2 shows a graphical representation of each of these metrics.

**Table 2.** Original dataset without balancing and optimization

| | Accuracy | Precision | Recall | F1-score | AUC-ROC |
|---|---|---|---|---|---|
| KNN | 72.73% | 66.67% | 56.82% | 61.35% | 69.67% |
| SVM | 77.49% | 84.62% | 50.00% | 62.86% | 72.20% |
| RF | 77.35% | 71.95% | 67.05% | 69.41% | 75.48% |
| ET | 78.36% | 75.68% | 63.64% | 69.14% | 75.52% |
| GB | 77.92% | 70.79% | 71.59% | 71.19% | 76.70% |

**Table 3.** Balanced dataset + t-SNE

| | Accuracy | Precision | Recall | F1-score | AUC-ROC |
|---|---|---|---|---|---|
| KNN | 77.50% | 75.93% | 81.19% | 78.47% | 77.46% |
| KNN+GS | 84.50% | 78.69% | 95.05% | 86.10% | 84.39% |
| SVM | 72.50% | 71.70% | 75.25% | 73.43% | 72.47% |
| SVM+GS | 90.50% | 93.62% | 87.13% | 90.26% | 90.53% |
| RF | 85.31% | 82.14% | 91.09% | 86.38% | 85.44% |
| RF+GS | 85.50% | 81.58% | 92.08% | 86.51% | 85.63% |
| ET | 87.50% | 83.33% | 94.06% | 88.37% | 87.43% |
| ET+GS | 88.50% | 84.82% | 94.06% | 89.20% | 88.44% |
| GB | 78.50% | 76.85% | 82.18% | 79.43% | 78.46% |
| GB+GS | 82.00% | 79.82% | 86.14% | 82.86% | 81.96% |

**Table 4.** Balanced dataset + MDS

| | Accuracy | Precision | Recall | F1-score | AUC-ROC |
|---|---|---|---|---|---|
| KNN | 76.00% | 73.87% | 81.19% | 77.36% | 75.95% |
| KNN+GS | 87.00% | 80.49% | 98.02% | 88.39% | 86.89% |
| SVM | 77.00% | 77.23% | 77.23% | 77.23% | 77.00% |
| SVM+GS | 84.00% | 82.86% | 86.14% | 84.47% | 83.98% |
| RF | 89.50% | 85.09% | 94.87% | 89.71% | 89.43% |
| RF+GS | 90.50% | 86.61% | 94.93% | 90.58% | 90.44% |
| ET | 91.50% | 88.89% | 94.17% | 91.45% | 91.46% |
| ET+GS | 92.00% | 89.72% | 95.05% | 92.31% | 91.97% |
| GB | 82.50% | 81.73% | 84.16% | 82.93% | 82.48% |
| GB+GS | 90.00% | 85.84% | 93.87% | 89.68% | 89.94% |

**Table 5.** Balanced dataset + KBEST

| | Accuracy | Precision | Recall | F1-score | AUC-ROC |
|---|---|---|---|---|---|
| KNN | 78.50% | 77.36% | 81.19% | 79.23% | 78.47% |
| KNN+GS | 86.50% | 80.33% | 97.03% | 87.89% | 86.39% |
| SVM | 80.50% | 78.18% | 85.15% | 81.52% | 80.45% |
| SVM+GS | 85.00% | 84.47% | 86.14% | 85.30% | 84.99% |
| RF | 90.50% | 86.61% | 96.04% | 91.08% | 90.44% |
| RF+GS | 91.00% | 86.73% | 96.03% | 91.14% | 90.94% |
| ET | 91.50% | 89.62% | 94.06% | 91.79% | 91.47% |
| ET+GS | 92.50% | 90.57% | 97.05% | 93.70% | 92.47% |
| GB | 87.00% | 83.78% | 92.08% | 87.73% | 86.95% |
| GB+GS | 90.00% | 85.22% | 97.03% | 90.74% | 89.93% |

**Table 6.** Balanced dataset + Variance threshold

| | Accuracy | Precision | Recall | F1-score | AUC-ROC |
|---|---|---|---|---|---|
| KNN | 75.50% | 74.53% | 78.22% | 76.33% | 75.47% |
| KNN+GS | 86.00% | 79.67% | 97.03% | 87.50% | 85.89% |
| SVM | 79.00% | 76.58% | 84.16% | 80.19% | 78.95% |
| SVM+GS | 81.50% | 79.09% | 86.14% | 82.46% | 81.45% |
| RF | 90.50% | 85.96% | 96.03% | 90.72% | 90.43% |
| RF+GS | 91.50% | 87.61% | 97.04% | 92.08% | 90.64% |
| ET | 90.00% | 87.16% | 94.06% | 90.48% | 89.96% |
| ET+GS | 90.50% | 86.61% | 96.04% | 91.08% | 90.44% |
| GB | 86.00% | 84.11% | 89.11% | 86.54% | 85.97% |
| GB+GS | 88.50% | 84.21% | 95.05% | 89.30% | 88.43% |

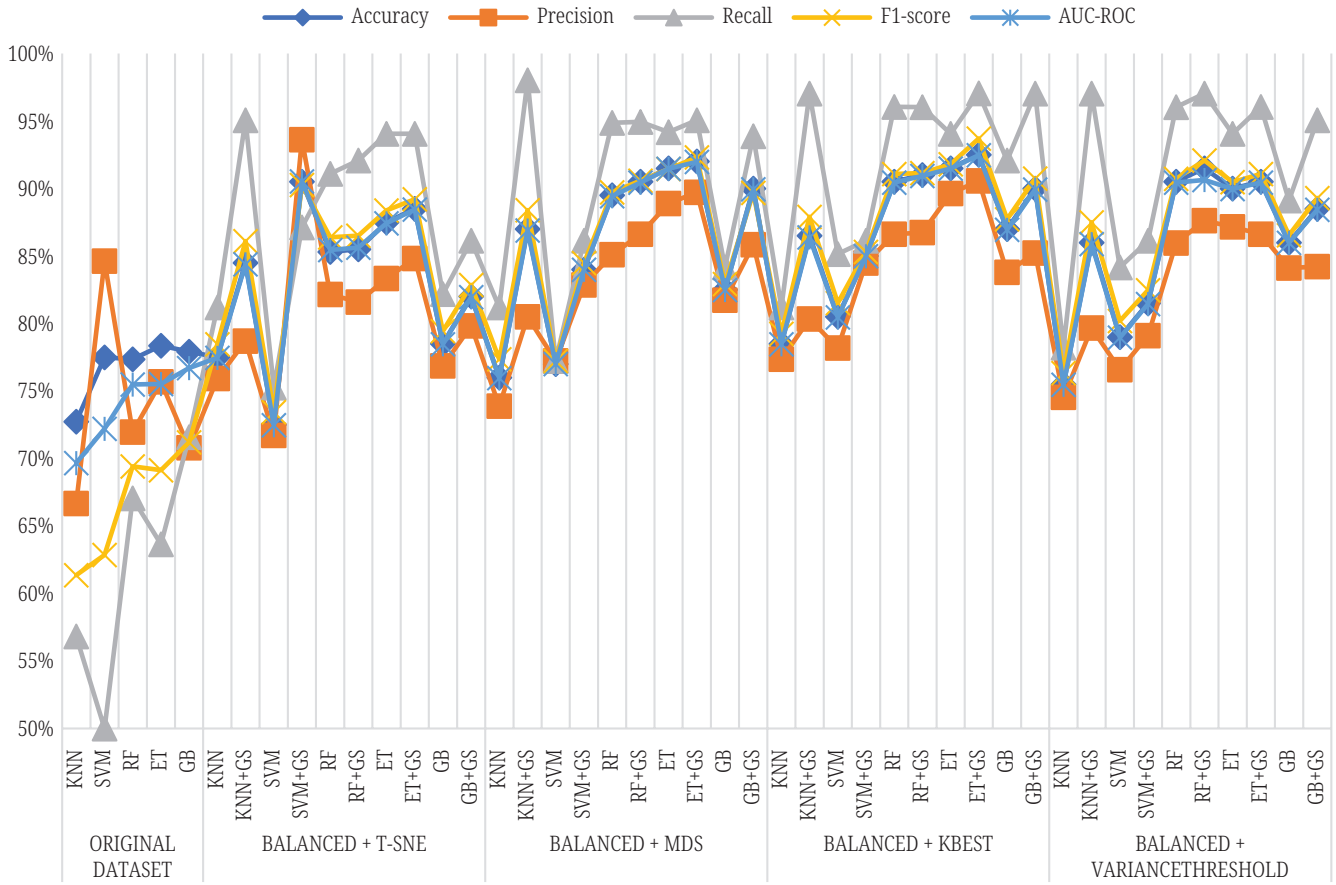


**Fig. 2.** Metrics of different algorithms

## 6 COMPARISON WITH RELATED WORK

Our approach emphasizes handling class imbalances and selecting relevant features, in agreement with previous research. The comparative analysis reveals

the superiority of our methodology for processing complex datasets and achieving high-quality predictive results. Numerous techniques have been employed in the existing literature for patient classification and diabetes prediction. Researchers frequently evaluate and compare the efficacy of various methods on a common dataset to identify optimal ones. The method choice depends on the dataset's specific characteristics and the research question. Key considerations include the model's interpretability and its practical applicability for clinical decision-making. Thus, the outcomes of our approach have been evaluated with those obtained by current approaches tailored for the same database, and the findings are summarized in Table 7.

**Table 7.** Related work comparison

| Authors | Methods | Best Algorithm Metrics |
|---|---|---|
| Al-Zebari et al. [13] | Logistic regression, decision trees, discriminant analysis, SVM, and KNN | Accuracy = 77.9% |
| Sonar et al. [14] | Decision Tree, ANN, Naive Bayes, and SVM | Precision = 85% |
| Kopitar et al. [15] | Glmnet, RF, XGBoost, LightGBM, and regression model | RMSE = 0.838 |
| García-Ordás et al. [16] | DNN, VAE, SAE, and CNN | Accuracy = 92.31% |
| Khanam et al. [17] | KNN, DT, AB, RF, SVM, NB, LR, and ANN | Accuracy = 88.6% |
| Ashraf Uddin et al. [18] | DT, RF, LR, and SVM | Accuracy = 80% |
| Febrian et al. [19] | KNN and Naive Bayes | Accuracy = 76.07%, Recall = 71.37%, Precision = 73.37% |
| Sihlangu et al. [20] | SGD, SVM, LR, and CN2 Rule | Accuracy = 80.7% |
| Mousa et al. [21] | LSTM, RF, and CNN | Accuracy = 85%, Precision = 82%, Recall = 78%, F1-score = 80%, AUC-ROC = 89% |
| This work | KNN, SVM, RF, ET, GB | Accuracy = 92.50%, Precision = 90.57%, Recall = 97.05%, F1-score = 93.70%, AUC-ROC = 92.47% |

## 7 CONCLUSION

Our study demonstrates the effectiveness of advanced machine learning techniques in diabetes prediction. Through meticulous optimization phases encompassing class imbalance handling, feature selection, dimensionality reduction, and hyperparameter tuning, we significantly improved the performance of classification algorithms. The top-performing model was obtained by employing the balanced dataset with the KBEST approach. Through optimization, Extra Trees achieved an accuracy of 92.50% and an F1-score of 93.70%, outperforming other techniques such as the balanced dataset with t-SNE (90.50%), MDS (92.00%), and variance threshold (91.50%). This underscores the superiority of Extra Trees in our experimental framework. Our methodology also outperformed several existing methods, showcasing its

potential for real-world applications and early patient management. Future research will focus on integrating metaheuristic techniques for hyperparameter optimization and validating our methodology with diverse datasets to enhance its generalizability and applicability in clinical settings.

## 8 REFERENCES


[1] J. L. Harding, M. E. Pavkov, D. J. Magliano, J. E. Shaw, and E. W. Gregg, "Global trends in diabetes complications: A review of current evidence," *Diabetologia*, vol. 62, no. 1, pp. 3–16, 2019. https://doi.org/10.1007/s00125-018-4711-2

[2] "Diabetes," Accessed: Feb. 21, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes.

[3] C. R. Whitehouse *et al.*, "Economic impact and health care utilization outcomes of diabetes self-management education and support interventions for persons with diabetes: A systematic review and recommendations for future research," *The Science of Diabetes Self-Management and Care*, vol. 47, no. 6, pp. 457–481, 2021. Accessed: Feb. 21, 2024. [Online]. Available: https://journals.sagepub.com/doi/abs/10.1177/26350106211047565.

[4] M. Ortiz-Martínez, M. González-González, A. J. Martagón, V. Hlavinka, R. C. Willson, and M. Rito-Palomares, "Recent developments in biomarkers for diagnosis and screening of type 2 diabetes mellitus," *Curr Diab Rep*, vol. 22, no. 3, pp. 95–115, 2022. https://doi.org/10.1007/s11892-022-01453-4

[5] American Diabetes Association, "2. Classification and diagnosis of diabetes: Standards of medical care in diabetes—2021," *Diabetes Care*, vol. 44, no. Supplement_1, pp. S15–S33, 2021. https://doi.org/10.2337/dc21-S002

[6] M. Shehab *et al.*, "Machine learning in medical applications: A review of state-of-the-art methods," *Computers in Biology and Medicine*, vol. 145, p. 105458, 2022. https://doi.org/10.1016/j.compbiomed.2022.105458

[7] M. S. Alzboon, M. S. Al-Batah, M. Alqaraleh, A. Abuashour, and A. F. H. Bader, "Early diagnosis of diabetes: A comparison of machine learning methods," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 15, pp. 144–165, 2023. https://doi.org/10.3991/ijoe.v19i15.42417

[8] N. Noorbakhsh-Sabet, R. Zand, Y. Zhang, and V. Abedi, "Artificial intelligence transforms the future of health care," *The American Journal of Medicine*, vol. 132, no. 7, pp. 795–801, 2019. https://doi.org/10.1016/j.amjmed.2019.01.017

[9] D. Cedeno-Moreno and M. Vargas-Lombardo, "Mobile applications for diabetes self-care and approach to machine learning," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 16, no. 8, pp. 25–38, 2020. https://doi.org/10.3991/ijoe.v16i08.13591

[10] Haviluddin, N. Puspitasari, A. E. Burhandeny, A. D. A. Nurulita, and D. Trahutomo, "Naïve Bayes and K-Nearest Neighbor algorithms performance comparison in diabetes mellitus early diagnosis," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 18, no. 15, pp. 202–215, 2022. https://doi.org/10.3991/ijoe.v18i15.34143

[11] J. Chaki, S. Thillai Ganesh, S. K. Cidham, and S. Ananda Theertan, "Machine learning and artificial intelligence based diabetes Mellitus detection and self-management: A systematic review," *Journal of King Saud University – Computer and Information Sciences*, vol. 34, no. 6, Part B, pp. 3204–3225, 2022. https://doi.org/10.1016/j.jksuci.2020.06.013

[12] N. Alanazi, Y. Alruwaili, A. Alazmi, A. Alazmi, M. Alanazi, and W. Alruwaili, "A systematic review of machine learning and artificial intelligence for diabetes care," *Journal of Health Informatics in Developing Countries*, vol. 17, no. 1, pp. 1–15, 2023. Accessed: Feb. 21, 2024. [Online]. Available: https://jhidc.org/index.php/jhidc/article/view/401.

[13] A. Al-Zebari and A. Sengur, "Performance comparison of machine learning techniques on diabetes disease detection," in *2019 1st International Informatics and Software Engineering Conference (UBMYK)*, Ankara, Turkey, 2019, pp. 1–4. https://doi.org/10.1109/UBMYK48245.2019.8965542

[14] P. Sonar and K. JayaMalini, "Diabetes prediction using different machine learning approaches," in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2019, pp. 367–371. https://doi.org/10.1109/ICCMC.2019.8819841

[15] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, "Early detection of type 2 diabetes mellitus using machine learning-based prediction models," *Sci Rep*, vol. 10, no. 1, 2020. https://doi.org/10.1038/s41598-020-68771-z

[16] M. T. García-Ordás, C. Benavides, J. A. Benítez-Andrades, H. Alaiz-Moretón, and I. García-Rodríguez, "Diabetes detection using deep learning techniques with oversampling and feature augmentation," *Computer Methods and Programs in Biomedicine*, vol. 202, p. 105968, 2021. https://doi.org/10.1016/j.cmpb.2021.105968

[17] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Express*, vol. 7, no. 4, pp. 432–439, 2021. https://doi.org/10.1016/j.icte.2021.02.004

[18] Md. A. Uddin *et al.*, "Machine learning based diabetes detection model for false negative reduction," *Biomedical Materials & Devices*, vol. 2, no. 1, pp. 427–443, 2023. https://doi.org/10.1007/s44174-023-00104-w

[19] M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryanigrum, and R. Yunanda, "Diabetes prediction using supervised machine learning," *Procedia Computer Science*, vol. 216, pp. 21–30, 2023. https://doi.org/10.1016/j.procs.2022.12.107

[20] N. Sihlangu and R. C. Millham, "Analysis of machine learning methods to determine the best data analysis method for diabetes prediction," in *2023 Conference on Information Communications Technology and Society (ICTAS)*, Durban, South Africa, 2023, pp. 1–6. https://doi.org/10.1109/ICTAS56421.2023.10082727

[21] A. Mousa, W. Mustafa, R. B. Marqas, and S. H. M. Mohammed, "A comparative study of diabetes detection using the Pima Indian diabetes database," *Journal of Duhok University*, vol. 26, no. 2, pp. 277–288, 2023. https://doi.org/10.26682/sjuod.2023.26.2.24

[22] "Pima Indians Diabetes Database – dataset by data-society," data.world. Accessed: Feb. 21, 2024. [Online]. Available: https://data.world/data-society/pima-indians-diabetes-database.

[23] P. Misra and A. S. Yadav, "Impact of preprocessing methods on healthcare predictions," in *Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE) 2019,* Rochester, NY, Mar. 09, 2019. https://doi.org/10.2139/ssrn.3349586

[24] M. Khushi *et al.*, "A comparative performance analysis of data resampling methods on imbalance medical data," *IEEE Access*, vol. 9, pp. 109960–109975, 2021. https://doi.org/10.1109/ACCESS.2021.3102399

[25] A. A. Aouragh and M. Bahaj, "Advanced cardiovascular disease diagnosis with machine learning: Exploring KBest, t-SNE, grid search, and ensemble methods," in *2023 IEEE International Conference on Advances in Data-Driven Analytics and Intelligent Systems (ADACIS)*, Marrakesh, Morocco, 2023, pp. 1–6. https://doi.org/10.1109/ADACIS59737.2023.10424089

[26] K. Balabaeva and S. Kovalchuk, "Comparison of efficiency, stability and interpretability of feature selection methods for multiclassification task on medical tabular data," in *Computational Science – ICCS 2021*, M. Paszynski, D. Kranzlmüller, V. V. Krzhizhanovskaya, J. J. Dongarra, and P. M. A. Sloot, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021, pp. 623–633. https://doi.org/10.1007/978-3-030-77967-2_51

[27] A. A. Aouragh, M. Bahaj, and N. Gherabi, "Comparative study of dimensionality reduction techniques and machine learning algorithms for Alzheimer's disease classification and prediction," in *2022 IEEE 3rd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, Fez, Morocco, 2022, pp. 1–6. https://doi.org/10.1109/ICECOCS55148.2022.9983211

[28] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN COMPUT. SCI.*, vol. 2, no. 3, p. 160, 2021. https://doi.org/10.1007/s42979-021-00592-x

[29] M. R. Camana Acosta, S. Ahmed, C. E. Garcia, and I. Koo, "Extremely randomized trees-based scheme for stealthy cyber-attack detection in smart grid networks," *IEEE Access*, vol. 8, pp. 19921–19933, 2020. https://doi.org/10.1109/ACCESS.2020.2968934

[30] B. Bischl *et al.*, "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 2, p. e1484, 2023. https://doi.org/10.1002/widm.1484

[31] M. Z. Naser and A. H. Alavi, "Error metrics and performance fitness indicators for artificial intelligence and machine learning in engineering and sciences," *Archit. Struct. Constr.*, vol. 3, no. 4, pp. 499–517, 2023. https://doi.org/10.1007/s44150-021-00015-8

## 9 AUTHORS

**Abd Allah Aouragh** is pursuing his Ph.D. at the MIET Laboratory, Faculty of Sciences and Techniques, Hassan First University, Settat, Morocco. He focuses on advancing medical diagnosis support systems. His research interests encompass machine learning, deep learning, and computer vision (E-mail: a.aouragh@uhp.ac.ma).

**Mohamed Bahaj** is a professor of computer sciences at the Faculty of Sciences and Techniques, Hassan First University, Settat, Morocco, where he conducts research in Artificial Intelligence, Software Engineering, and Data Mining.

**Fouad Toufik** is a professor of computer sciences at the Higher School of Technology SALE, Mohammed V University, Morocco. His research interests focus on artificial intelligence, big data, and database architectures.