Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 8, S. 284-295, 5, 2024

Araştırma Makalesi

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 8, pp. 284-295, 5, 2024

Research Article

https://as-proceeding.com/index.php/ijanser
ISSN: 2980-0811

Enhancing Diabetes Diagnosis through the Investigation of Cost-Sensitive Learning with Ensemble Techniques

Yasmine Khedimi*, Nadjette Dendani[2] and Hana Khemisa[3]

[1, 2,3]*Dept. of Computer Science, Badji Mokhtar University, Annaba, Algeria*

yasminekhedimi893@gmail.com

(Received: 24 June 2024, Accepted: 27 June 2024)

(3rd International Conference on Frontiers in Academic Research ICFAR 2024, June 15-16, 2024)

ATIF/REFERENCE: Khedimi, Y., Dendani, N. & Khemisa, H. (2024). Enhancing Diabetes Diagnosis through the Investigation of Cost-Sensitive Learning with Ensemble Techniques. *International Journal of Advanced Natural Sciences and Engineering Researches*, 8(5), 284-295.

Abstract – Diabetes mellitus is a prevalent chronic disease represented in the body's un-successful insulin effect, that appears in the elevation of the blood's glucose levels and potential damage to many body systems, causing the increasing of mortality rates. Early diagnosis is important for managing this illness, and machine learning algorithms play a crucial role employing various methodologies for diabetes detection, and in handling imbalanced data in particular.
Using diverse classification algorithms such as (Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Convolutional Neural Network) for diabetes diagnosis and classification demonstrate the dominance of one class and the resulting underrepresentation of the minority class.
To address this issue, cost-sensitive learning and resampling techniques are investigated in this study. The proposed approach aimed to propose robust cost-sensitive classifiers by modifying the objective functions of well-known algorithms. Additionally, hybrid approach of our improved Cost-sensitive models with well used ensemble techniques like Cost-sensitive XGBoost and Cost-sensitive Random Forest, Cost-sensitive Logistic Regression are analyzed to effectively address imbalanced classes.
To validate proposed models two imbalanced medical datasets (PIMA Indi-an, and BASEDIABET datasets) are applied. Obtained results proves the accuracy and sensitivity of diabetes prediction models enhancement, by reducing costly classification errors.

Keywords – Classification problem, Imbalanced datasets, Algorithm-level solutions, Ensemble techniques, Cost-sensitive learning, diabetes mellitus diagnosis

I. INTRODUCTION

In the last few years, diabetes-related diseases have emerged as a leading cause of death in the developing world. The interplay of genetics and lifestyle significantly influences personalized treatment, causing an elevation in blood sugar levels and increasing long-term health risks' possibilities [1]. Current research is focused on advancing technologies such as continuous glucose monitoring (CGM) [2] and artificial pancreas development to uncover novel biomarkers and diagnostic methodologies for early diabetes detection, enhancing our understanding of this condition. The field of machine learning is becoming increasingly critical in the realm of artificial intelligence, as it employs algorithms trained on

data to develop adaptable models capable of handling complex tasks. In the medical field, especially concerning diabetes detection, machine learning holds considerable promise

However, learning from imbalanced datasets, where one class is dominant and another is underrepresented, presents significant challenges. Traditional classifiers face difficulties with such imbalances, resulting in skewed datasets and struggles to accurately identify rare cases. This scenario can lead to the creation of misleading models and difficulties in distinguishing between small, overlapping classes. To address these challenges, several approaches can be employed, including solutions at the algorithm level. Cost-sensitive learning is one such algorithm-level solution, involving the modification of algorithms or their objective functions to consider the costs associated with misclassification errors [3]. This modification allows the algorithm to prioritize correctly classifying the minority class, which is often more costly to misclassify than the majority class. Another algorithm-level solution is the use of ensemble techniques, which focus on balancing the class distribution to improve model performance. Other algorithm-level solutions include modifying classifier methods or optimizing the performance of learning algorithms.

In this study, we conducted a comprehensive investigation by individually applying various ensemble techniques, including RUSBoost (Random Under Sampling Boosting) and Balanced Random Forest and XGBoost (eXtreme Gradient Boosting). Additionally, we explored cost-sensitive learning, an approach used in machine learning to consider the costs associated with classification errors. The experimentation encompassed the utilization of 5 biased classifiers, namely Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Convolutional Neural Network (CNN). Our objective was to assess the performance of each technique on these classifiers in the context of addressing class imbalance. Subsequently, we proposed the hybridization of some of these techniques (Cost-sensitive XGBoost, Cost-sensitive Random Forest, Cost-sensitive Decision Tree), aiming to uncover the most effective method for mitigating the challenges associated with imbalanced classes. The study contributes valuable insights into the nuanced dynamics of these methods and their combined effectiveness in enhancing classification outcomes.

Furthermore, we have improved the hyperparameters of the cost-sensitive XGBoost, cost-sensitive Random Forest, and cost-sensitive CNN to further enhance their performance and achieve more accurate classifications. The empirical investigation was conducted using two datasets related to diabetes disease, including the PIMA Indian data diabetes dataset [4] and the DataBase-Diabetes database designed in 2018 [5]. These datasets hold considerable prominence in diabetes research, with recent studies showcasing positive outcomes.

Our study aims to enhance the accuracy of diabetes diagnosis by addressing class imbalance and improving classifier performance. By combining ensemble techniques and cost-sensitive learning, we aim to create robust models that can effectively handle imbalanced medical data. This hybrid approach leverages the strengths of both methods, resulting in more accurate predictions and overall better performance.

II. Related Work

Several studies have been conducted to develop tools for diagnosing diabetes.

In 2006, the study “Performance analysis of cost-sensitive learning methods with application to imbalanced medical data” delves deep into the use of cost-sensitive learning for imbalanced medical data classification, showcasing its advantages over traditional approaches and proposing novel solutions to enhance model performance in this specific context [6].

In 2014, the article “Learning to Improve Medical Decision Making from ImbalancedData without a Priori Cost” introduces the RankCost boosting algorithm based on cost-sensitive learning for predicting imbalanced medical data, maximizing the difference between majority and minority classes [7].

In 2020, the article “Cost-Sensitive Classification Algorithm for Imbalanced Data in Medical Diagnosis” integrated a naive Bayes algorithm augmented by a tree and the

AdaCost cost-sensitive algorithm to handle imbalanced medical data, achieving superior performance compared to some state-of-the-art methods [8].

In 2018, the article “Predicting Hospital Readmission via Cost-Sensitive Deep Learning” proposed a cost-sensitive deep learning approach to predict hospital

readmission, combining convolutional neural networks with a cost-sensitive MLP classifier to address class imbalance during model training [9].

Finally, the article "Ibomoiye Domor Mienye, Yanxia Sun, Performance analysis of cost-sensitive learning, methods with application to imbalanced medical data, Informatics in Medicine Unlocked, Volume 25, 2021,100690, ISSN 2352-9148 “[10].

Our study aims to enhance the accuracy of diabetes diagnosis by addressing class imbalance and improving classifier performance. By combining ensemble techniques and cost-sensitive learning, the researchers, aim to create robust models that can effectively handle imbalanced medical data. This hybrid approach leverages the strengths of both methods, resulting in more accurate predictions and overall better performance.

III. MATERIALS AND METHOD

In our innovative approach, we employed various machine learning models focuses on minimizing false negatives, even if it leads to a slight increase in false positives, to prioritize the identification of in-Divi duals with the disease. While logistic regression and other models may face challenges with imbalanced data, techniques like cost-sensitive learning, as employed in this study, can mitigate this issue by assigning higher weights to the minority class and generating syn-thetic data points. Implementing these techniques enhances model performance for medical diagnosis tasks with imbalanced data and varying misclassification costs, resulting in more accurate predictions.

A. Datasets and Assessment Criteria

In biomedical domains, it is common to encounter limited data representing less prevalent cases, while specialized domain knowledge is often readily available. For example, in diabetes databases, minority data might include rare instances of severe complications or uncommon disease variations, while domain expertise encompasses detailed information on risk factors, comorbidities, and optimal treatment protocols. This research utilizes two primary datasets: the Pima Indian Diabetes Database (PIDD) [4] and the BASEDIABET dataset. The PIDD, a widely used resource in diabetes studies, contains demographic and medical records from the Pima Native American community, including data points such as age, pregnancy history, blood glucose levels, and blood pressure. The Base-diabete dataset, established in 2022, includes authentic samples with variables such as age, height, weight, body mass index (BMI), glycosylated hemoglobin (HbA1c), and diabetes type, comprising 251 entries. The Base-diabete dataset was compiled by Rayane Allouani in 2018 at CHU Ibn Sina Hospital in Annaba, and from the offices of Dr. Bouali, Dr. Benaissa, Dr. Amraoui, and the Didouche Mourad polyclinic [5]. After preprocessing the databases by eliminating missing values and performing normalization, we partitioned the dataset into 80% for training and 20% for testing due to the limited number of samples.

Table 1. Descriptive table of the datasets.

Datasets	Features	Samples	Majority	Minority
Pima	9	768	500	268
BASE-DIABETE	7	251	151	99

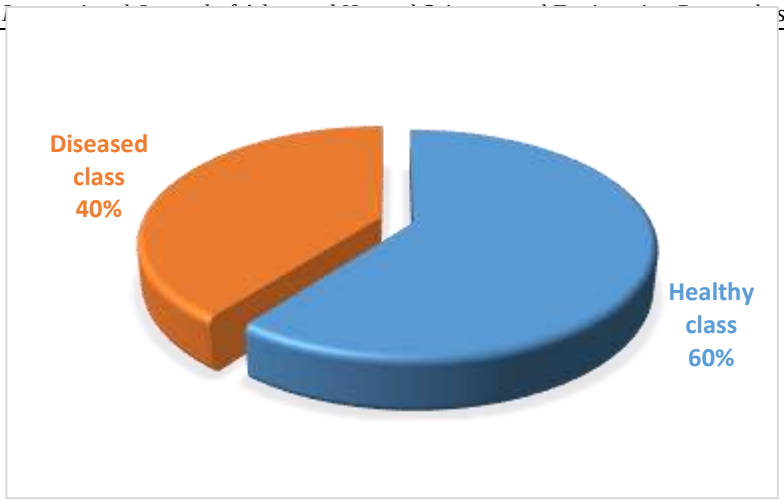

Fig 1. Base-Diabete imbalance Issue

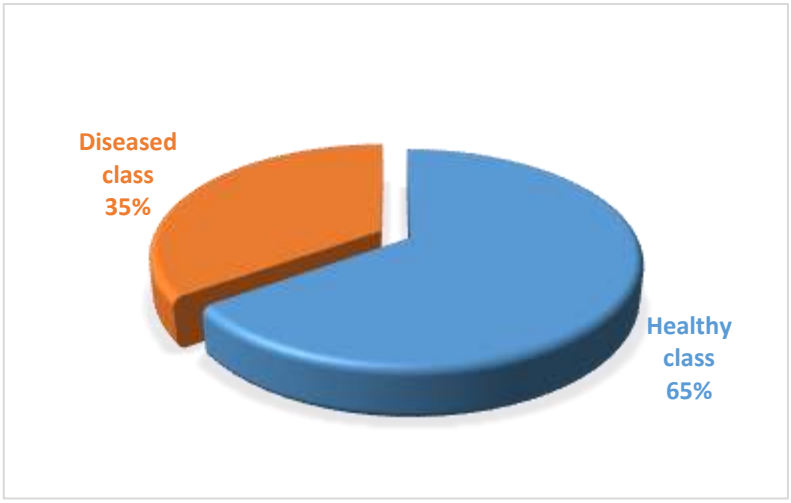

Fig2. Pima Imbalance Issue

B. *Methods*

Here, we explore some methods utilized and put into practice in this research.

- *RUSBoost* (Random Undersampling Boosting) combines random undersampling of the majority class with the boosting technique to enhance the performance of a classifier on imbalanced datasets. It operates similarly to SMOTEBoost, but with a distinct approach to addressing class imbalance. Instead of creating synthetic instances for the minority class, RUSBoost focuses on the majority class by undersampling it randomly at each iteration. This means that examples from the majority class are removed to balance the classes. Contrary to SMOTEBoost, RUSBoost does not need to assign new weights to the instances. It simply normalizes the weights of the remaining instances in the new dataset relative to their total weight sum. The rest of the procedure follows the same steps as in SMOTEBoost, where new classifiers are trained on weighted datasets, and their performance is evaluated on the validation dataset. However, it's worth noting that cost-sensitive learning cannot be directly added to RUSBoost because it already considers the cost by undersampling the majority class [11].

- *Balanced Random Forest* is a variant of Random Forests that is originally designed to minimize errors. Two bootstrap ensembles of the same size are constructed: one for the minority class and one for the majority class. These two ensembles together form the training set.

Afterward, the Random Forest algorithm is applied as usual. A bootstrap ensemble is a subset created by repeatedly sampling instances with replacement from an original dataset [12].

- *Cost-sensitive learning* in binary classification involves minimizing the empirical risk by Regularized Empirical Risk Minimization (ERM). The empirical risk (Remp) is calculated as the sum of the loss function (L) applied to the model's predictions f(xi) and the true labels (yi) for each data point in the training set, divided by the total number of data points N:

$$Remp = \frac{1}{N} \sum_{I=1}^{N} L(f(xi), yi) \quad \textbf{(1)}$$

Minimizing the empirical risk helps the model perform well on the training data, but it can lead to overfitting. To prevent overfitting, regularization is applied, adding a penalty term Ω(f) to the objective function. This discourages overly complex models in favor of simpler ones that are less likely to overfit. The combined objective function is:

$$R = Remp + \lambda \cdot \Omega(\mathrm{f}) \quad \textbf{(2)}$$

Here, λ is a hyperparameter that balances minimizing the empirical risk and controlling model complexity. This approach helps the model generalize better to unseen data, improving its overall performance.

Cost-sensitive learning recognizes the real-world consequences of errors and assigns different weights or costs to different types of mis classification based on their impact. In medical diagnosis, where imbalanced datasets are common, such as in diabetes prediction, the concept of variable misclassification costs is crucial. For example, false negatives (missing a case of diabetes) can be more costly than false positives (incorrectly predicting diabetes) due to potential treatment delays [13].

- *Cost-Sensitive Convolutional Neural Network (CNN)*) is a variant of traditional CNNs that incorporates the cost associated with classification errors during the model training process. This approach allows the model to prioritize correcting the costliest errors, which is particularly useful in scenarios where certain mistakes have more severe consequences than others. Cost-Sensitive CNNs use a modified loss function those weights classification errors according to their respective costs. This adjusted loss function directs the learning process, encouraging the model to focus on rectifying the most expensive mistakes. The benefits of Cost-Sensitive CNNs include improved overall performance by focusing on the costliest errors, reduction of actual costs such as financial losses or physical damage by minimizing the most expensive errors, and better adaptation to specific problems by allowing for the customization of personalized costs to different types of errors. Cost-Sensitive CNNs are particularly useful in fields like medical detection, where misclassification can have severe implications for a patient's health, enhancing diagnostic accuracy by focusing more on correcting the most critical errors. In our study, we have improved the hyperparameters of the cost-sensitive CNN to further enhance its performance [14].

- *Cost-sensitive XGBoost* is an ensemble algorithm that aggregates machine learning trees using the gradient boosting principle. It is widely used in various classification and regression tasks, demonstrating strong performance [4], particularly in scenarios with imbalanced class distributions. To further enhance its effectiveness, we introduce a modification known as cost-sensitive XGBoost. This variant focuses more on correctly classifying the minority class by adjusting the algorithm's behavior during training. In scikit-learn, this adjustment is achieved through a hyperparameter called scale_pos_weight. By default, this parameter is set to 1.0 in XGBoost. However, to improve performance, we can set it to the inverse of the class

distribution. This scaling factor influences how the algorithm treats errors in the minority class during training, encouraging it to prioritize correcting these errors. Additionally, we have improved the hyperparameters of the cost-sensitive XGBoost to further enhance its classification capabilities. As a result, the model achieves better performance when classifying instances from the minority class [15].

- *Cost-sensitive decision tree* is tailored for imbalanced classification tasks, a scenario where traditional decision tree algorithms may falter due to their bias towards majority class instances. Unlike traditional approaches that prioritize sample separation without considering minority class importance, cost-sensitive decision trees adjust their split point selection to give priority to minority class instances, effectively addressing the class imbalance. This adjustment is achieved by computing purity using metrics such as the Gini index or entropy, taking into account the class distribution within a group. For example, in CART implementations, the Gini index is typically used for purity computation. The algorithm modifies the splitting criteria by assigning higher weights to minority class instances and lower weights to majority class instances, which is often based on the inverse of class distribution in the dataset. This prioritization ensures that the algorithm focuses on minority class instances during node purity calculations, leading to enhanced performance on imbalanced datasets [16].

- *Cost-sensitive random forest* is a variant of the random forest ensemble learning algorithm, designed to address imbalanced classification tasks. It constructs multiple decision trees during training and outputs the mode of classes for classification or the mean prediction for regression. While random forest helps mitigate overfitting, its performance can be influenced by dataset characteristics. To improve its performance on imbalanced datasets, class weights are introduced, encouraging the algorithm to prioritize the minority class by assigning weights based on the inverse of class distribution. This adjustment enhances random forest's ability to handle imbalanced data and improves classification accuracy in such scenarios. Additionally, hyperparameters of the cost-sensitive Random Forest have been improved to further optimize its performance on imbalanced datasets [17].

- *Cost-sensitive logistic regression* Standard logistic regression assumes a balanced class distribution, which may not be suitable for imbalanced datasets. To address this issue, a class weighting mechanism is introduced in logistic regression. This mechanism adjusts how the algorithm updates its coefficients during training, penalizing the model more for errors made on minority class samples and less for errors on majority class samples.

In standard logistic regression, the log-likelihood function (L(w)) is expressed as:

$$L(w) = \frac{1}{N}\sum_{i=1}^{N}[yi \ln\big(P(yi)\big) + (1 - yi)\ln\big(1 - P(yi)\big)] \quad \textbf{(3)}$$

Where P(yi) denotes the predicted probability that (y) is true for sample i.

In cost-sensitive logistic regression, the modified log likelihood function is represented:

$$L(w) = \frac{1}{N}\sum_{i=1}^{N}\big[C_{FP}\ yi \ln\big(P(yi)\big) + C_{FN}(1 - yi)\ln\,(1 - P(yi))\,\big] \quad \textbf{(4)}$$

This modification leads to a type of logistic regression that is well suited for imbalanced classification problems, known as cost-sensitive logistic regression [18].

C. Measures of performance

In our research, we employ several assessment metrics, including accuracy, precision, recall, F-measure, and Cohen's kappa coefficient, to assess the performance of our model. These metrics are

derived from the confusion matrix, where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively.

TP and TN represent the number of correct positive and negative predictions, while FP indicates instances where the model incorrectly predicts a healthy patient as sick, and FN denotes instances where the model fails to predict the presence of a disease when it is present.

The mathematical representations of these assessment metrics are as follows:

Accuracy: Accuracy is the proportion of correct predictions (both true positives and true negatives) out of the total number of predictions. It is calculated as:

$$Accuracy = \frac{TP+TN}{TP+TN-FP-FN} \quad \textbf{(5)}$$

Precision: Precision measures the proportion of true positive predictions out of all positive predictions made by the model. It indicates how many of the predicted positive cases are actually positive. It is calculated as:

$$P = \frac{TP}{TP+FP} \quad \textbf{(6)}$$

Recall: Recall, also known as sensitivity or true positive rate, measures the proportion of true positive predictions out of all actual positive cases. It indicates how well the model identifies positive cases. It is calculated as:

$$R = \frac{TP}{TP+FN} \quad \textbf{(7)}$$

F1-score: The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both. It is particularly useful when dealing with imbalanced datasets. It is calculated as:

$$F1 = 2 \times \frac{P \times R}{P+R} \quad \textbf{(8)}$$

Cohen's Kappa Coefficient: Cohen's kappa coefficient measures the agreement between the model's predictions and the actual outcomes, adjusting for the agreement that could occur by chance. It is calculated as:

$$\text{KPPA} = \frac{(P0-Pc)}{(1-Pc)} \quad \textbf{(9)}$$

where P0P_0P0 is the observed agreement and PcP_cPc is the expected agreement by chance.
These metrics provide a comprehensive evaluation of the model's performance, particularly in handling imbalanced datasets and minimizing misclassification errors.

Area Under the Receiver Operating Characteristic Curve (AUC-ROC): is a metric that measures a classifier's ability to distinguish between classes at various threshold levels. It represents the area under the ROC curve, which plots the true positive rate (TPR) against the false positive rate (FPR). The AUC-ROC score ranges from 0 to 1, with higher values indicating superior classifier performance in differentiating between classes.

$$AUC - ROC = \int_0^1 TPR(FPR)dFPR \quad \textbf{(10)}$$

IV. RESULTS

The following tables and figures present the results of our experiments evaluating the performance of various machine learning techniques for diabetes diagnosis classification. Through rigorous experimentation and analysis, these results highlight the effectiveness of different approaches.

Table 1. Performance assessment of the algorithms on the PIDD dataset.

Algorithm	Score F1	Accuracy	Precision	Recall	Kappa
SVM	0.601	0.733	0.645	0.563	0.403
CSSVM	0.687	0.740	0.602	0.8	0.472
XgBoost	0.649	0.733	0.612	0.690	0.436
CSXGBOOST	0.746	0.792	0.758	0.734	0.570
DT	0.608	0.707	0.583	0.636	0.453
CSDT	0.644	0.720	0.590	0.709	0.417
CNN	0.644	0.720	0.590	0.709	0.417
CSC NN	0.716	0.753	0.607	0.872	0.510
RF	0.583	0.740	0.65	0.528	0.397
CSRF	0.697	0.785	0.844	0.660	0.518
LR	0.647	0.759	0.68	0.618	0.465
CSLR	0.65	0.727	0.6	0.709	0.429
RUSBoost	0.651	0.707	0.567	0.763	0.409
Balanced Random Forest	0.673	0.740	0.601	0.765	0.463

Table 2. Performance assessment of the algorithms on the Base-Diabete dataset.

Algorithm	Score F1	Accuracy	Precision	Recall	Kappa
SVM	0.944	0.92	0.944	0.944	0.801
CSSVM	0.944	0.92	0.944	0.944	0.801
XgBoost	0.958	0.94	0.945	0.972	0.847
CSXGBOOST	0.969	0.96	1.0	0.941	0.911
DT	0.898	0.86	0.93	0.861	0.674
CSDT	0.857	0.8	0.882	0.833	0.524
CNN	0.857	0.8	0.882	0.833	0.524
CSC NN	0.914	0.88	0.941	0.888	0.714
RF	0.96	0.94	0.972	0.947	0.840
CSRF	0.866	0.84	0.764	0.98	0.675
LR	0.873	0.82	0.885	0.861	0.563
CSLR	0.868	0.8	0.825	0.916	0.456
RUSBoost	0.916	0.88	0.916	0.916	0.702

Balanced Random Forest	0.934	0.92	0.914	0.955	0.831

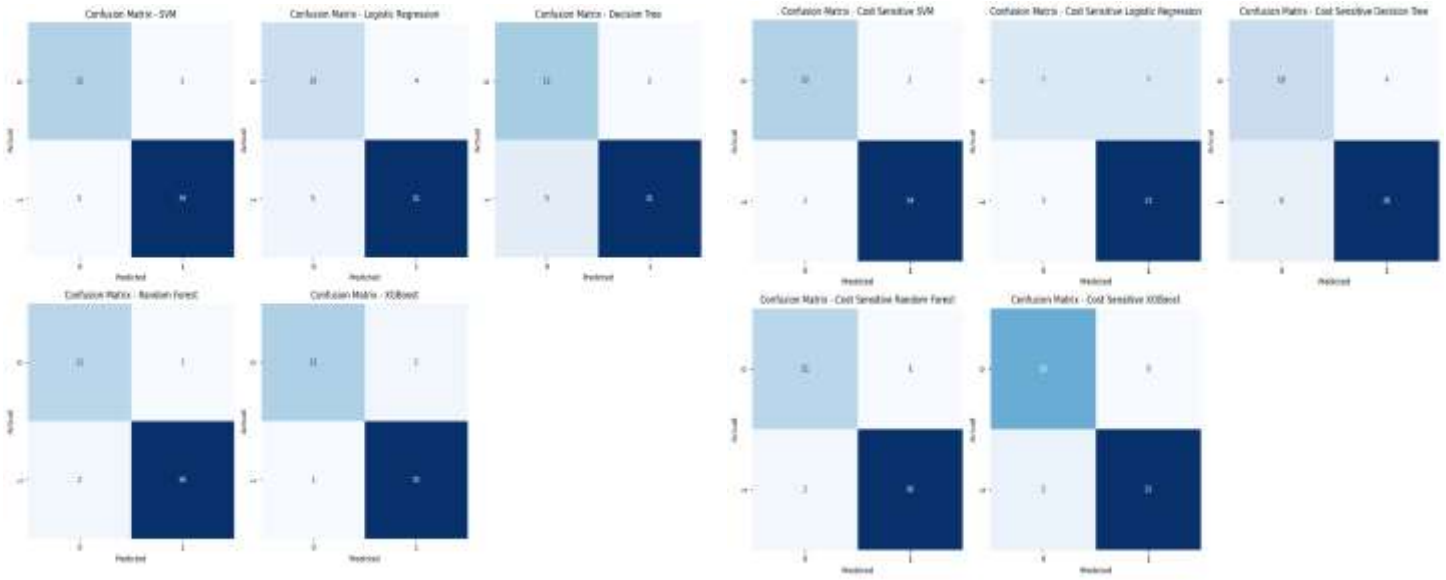

Fig 3. Confusion Matrices of Classifiers with and without Cost-Sensitive Learning for Base-Diabete dataset

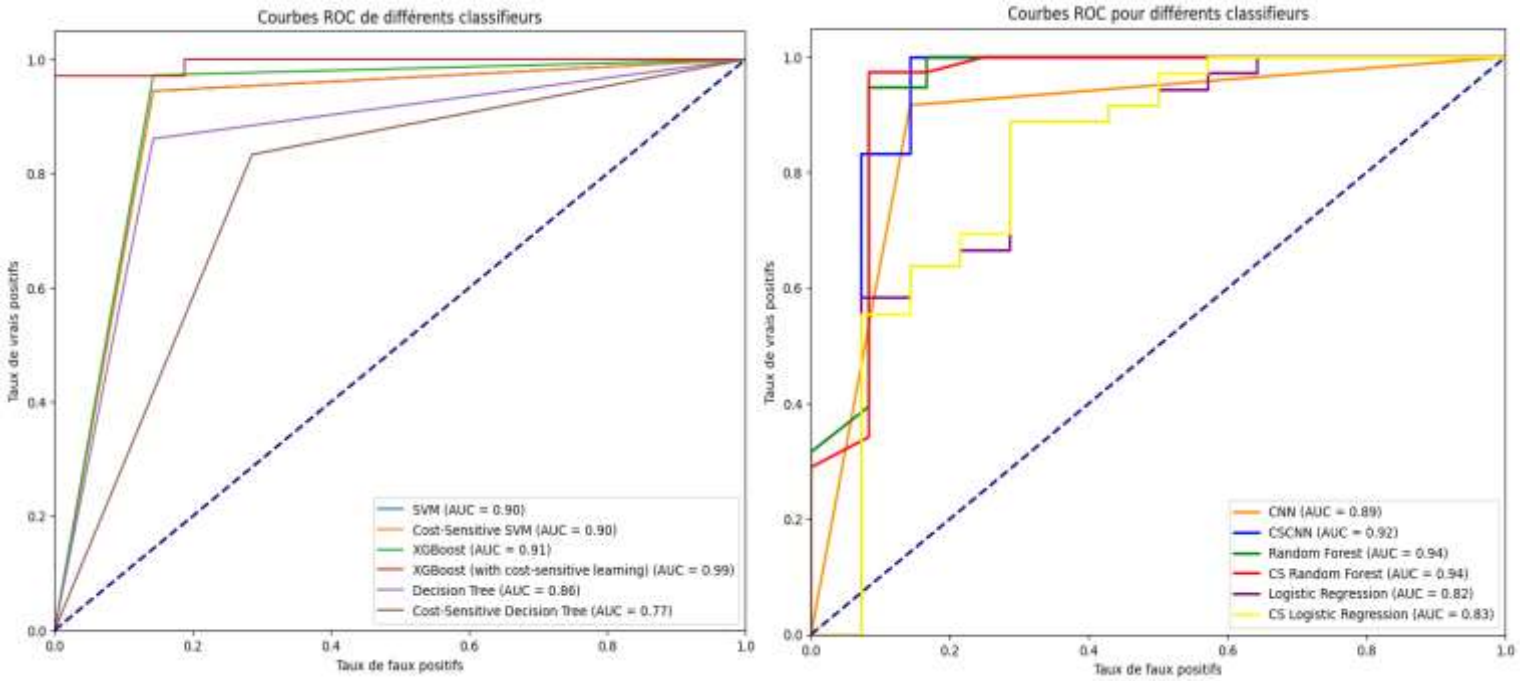

Fig4. Visualization of the AUC-ROC for different classifiers, both with and without cost-sensitive learning, on the Base-Diabete dataset.

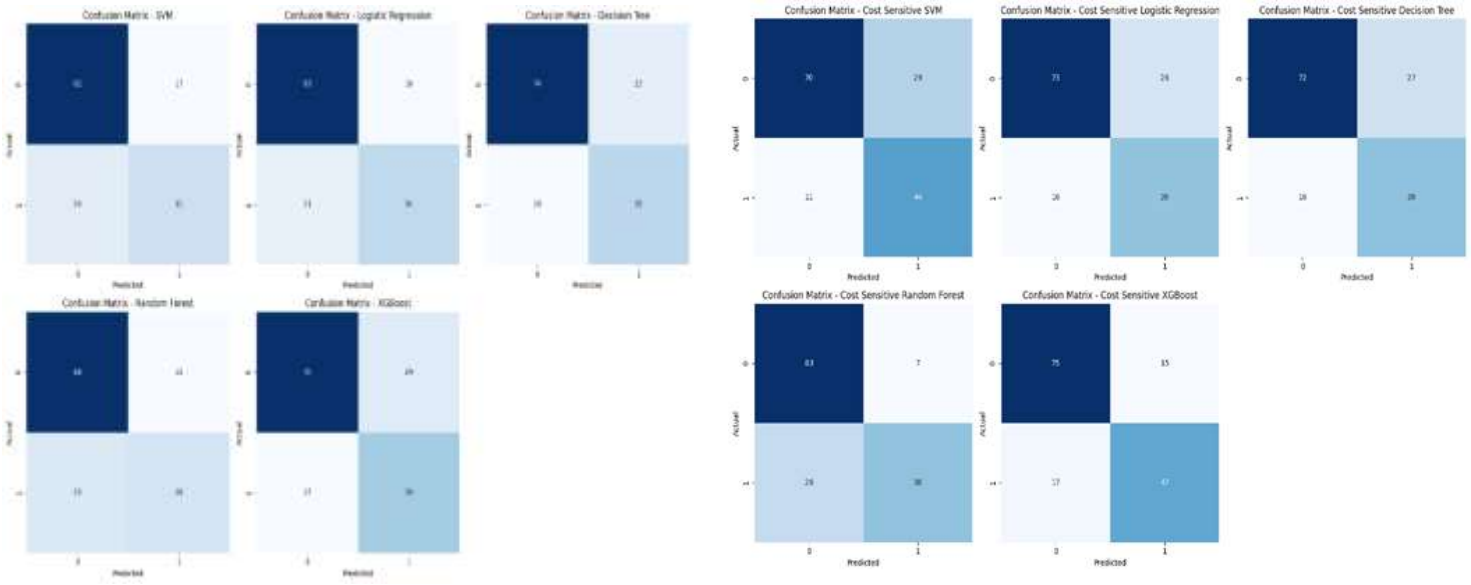

Fig5. Confusion Matrices of Classifiers with and without Cost-Sensitive Learning for PIDD

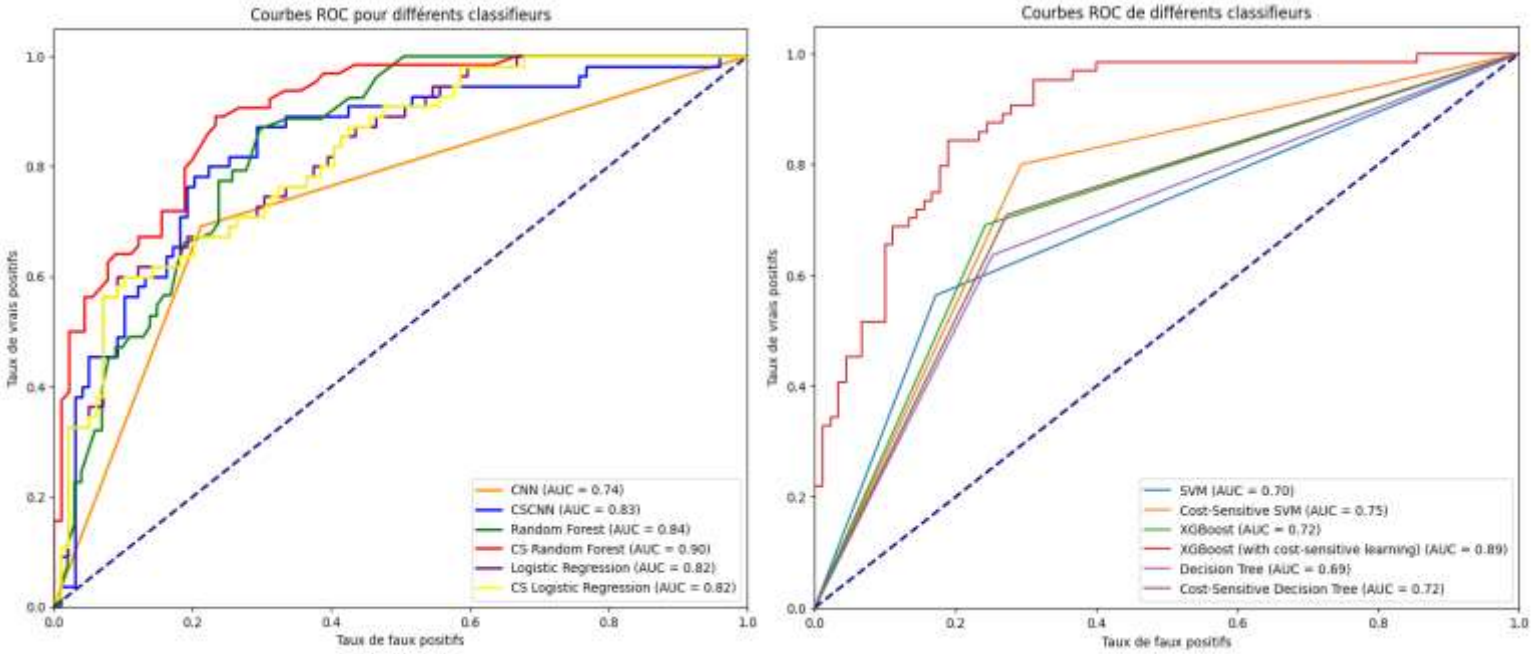

Fig4. Visualization of the AUC-ROC for different classifiers, both with and without cost-sensitive learning, on the PIDD dataset.

V. DISCUSSION

Analysis of Two Tables for Both Datasets:

In analyzing Table 1 for the PIDD dataset, CSSVM, CSXGBOOST, and CSCNN show higher scores in F1, Accuracy, Precision, Recall, and Kappa compared to their non-cost-sensitive counterparts. RUSBoost and Balanced Random Forest also demonstrate competitive performance, especially in Recall. The improved and modified parameters of Cost-Sensitive Random Forest, Cost-Sensitive CNN, and Cost-Sensitive XGBoost likely contributed to their enhanced performance, reflected in their higher scores. The

top two performing algorithms on the PIDD dataset based on accuracy are CSXGBOOST (0.792) and CSRF (0.785).

Similarly, for Table 2 on the Base-Diabet dataset, the enhanced parameters of Cost-Sensitive Random Forest, Cost-Sensitive CNN, and Cost-Sensitive XGBoost contribute to improved performance. CSXGBOOST and RF exhibit the best results, while RUSBoost and Balanced Random Forest also perform well with high recall values. Overall, CSXGBOOST and RF stand out as the best-performing models on the Base-Diabet dataset.

Analysis of confusion matrices for Both Datasets:

The analysis of confusion matrices for both the PIDD and Base-Diabet datasets indicates that cost-sensitive learning significantly enhances model performance. For the Base-Diabet dataset, Cost-Sensitive XGBoost is the most effective, maximizing true positives (increasing from 12 to 18) and eliminating false negatives. In contrast, other classifiers did not exhibit such notable improvements with the cost-sensitive approach. On the PIDD dataset, Random Forest excels in achieving a high number of true positives (TP=37), while XGBoost is best at minimizing false positives (FP=24). Despite Random Forest's balanced performance between true positives and false positives, there is a need for caution regarding false negatives. The analysis reveals that the optimal classifier for cost management varies by criterion: Random Forest is best for maximizing true positives, Decision Tree for minimizing false positives, and XGBoost for balancing true positives and false positives. The comparison between results with and without cost-sensitive learning shows diverse improvements across models, highlighting the importance of adapting the method to the specific needs of the classification task.

Analysis of AUC-ROC Curves for Both Datasets:

The visualization of the AUC-ROC for different classifiers for la base PIDD presents a ranking of classifiers with and without Cost-Sensitive (CS) adjustments, highlighting the differing performances of each model. Without CS adjustments, the Random Forest classifier achieves an AUC of 0.84, demonstrating excellent classification capability, followed by Logistic Regression with an AUC of 0.82, XGBoost with an AUC of 0.72, CNN with an AUC of 0.74, SVM with an AUC of 0.70, and Decision Tree with an AUC of 0.69. With CS adjustments, the CS Random Forest classifier, with modified parameters, achieves an exceptional AUC of 0.90, placing it at the top, followed by XGBoost with an AUC of 0.89, the CSCNN with an AUC of 0.83, CS Logistic Regression maintaining an AUC of 0.82, CS SVM with an AUC of 0.75, and CS Decision Tree with an AUC of 0.72. Overall, methods with cost-sensitive adjustments generally show better performance in terms of AUC compared to methods without CS adjustments, with XGBoost showing the most significant improvement (+0.17), likely due to parameter modifications for the cost-sensitive classifiers.

For the Base-Diabet dataset, XGBoost also showed a remarkable improvement with cost-sensitive learning, with its AUC increasing from 0.91 to 0.99, indicating that parameter modifications significantly enhanced its performance. Both CNN and Random Forest showed improvements, with CNN's AUC increasing from 0.89 to 0.92 and Random Forest's AUC increasing from 0.94 to 0.95, suggesting that parameter modifications had a positive impact. However, Decision Tree did not show better results with the cost-sensitive approach, with its AUC decreasing from 0.86 to 0.77, indicating that further parameter adjustment is needed or that this method is less suitable for this model in this context. Logistic Regression showed a minor improvement, with its AUC increasing from 0.82 to 0.83, indicating that it slightly benefits from the cost-sensitive approach but may require more tuning for significant gains.

VI. CONCLUSION

In conclusion, this study highlights the importance of machine learning algorithms in early diabetes detection and recognizes the challenges posed by imbalanced datasets in medical research. Researchers have developed robust cost-sensitive classifiers and integrated them with ensemble methods such as Cost-

sensitive XGBoost, Cost-sensitive Random Forest, and Cost-sensitive CNN, im-proving their hyperparameters to enhance performance and achieve more accurate results. This integrated approach shows promise in refining the classification of imbalanced medical data, thereby enhancing diabetes mellitus diagnostic methods. The study underscores the significance of innovative techniques for managing imbalanced datasets in medicine, offering potential for more precise and dependable predictions in healthcare. Future research could explore enriching datasets, investigating additional machine learning techniques such as pre-trained and transfer learning models, and incorporating real-time data for more dynamic insights into diabetes diagnosis. Pre-trained models, especially those trained on extensive healthcare datasets, could be optimized for diabetes diagnosis, potentially improving the efficiency and accuracy of the classification pro-cess. Additionally, transfer learning could be examined to leverage existing models for diabetes diagnosis, overcoming challenges related to limited data availa-bility and enhancing the models' generalizability across various healthcare set-tings.

REFERENCES

[1] World Health Organization. (2016). Global report on diabetes. [Online]. Available: https://www.who.int/publications/i/item/9789241565257.

[2] M. N. Reiley et al., "COVID-19 mRNA vaccines are immunogenic in cancer patients," JCI Insight, vol. 6, no. 6, Mar. 2021. [Online]. Available : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7755046/.

[3] M. Koziarski and K. Pancerz, "Stop Oversampling for Class Imbalance Learning: A Review," arXiv preprint arXiv:2202.03579, 2022. [Online]. Available: https://arxiv.org/pdf/2202.03579.

[4] Pima Indians Diabetes Database. [Online]. Available: https://kaggle.com/uciml/pima-indians-diabetes-database. [Accessed: 17-Apr-2021].

[5] R. Allouani and N. Dendani, "Design and Creation of an Expert System Based on an Ontology for Diagnosis of Diabetes," License thesis, Department of Computer Science, Badji Mokhtar Annaba University, Algeria, June 2018.

[6] Ibomoiye Domor Mienye, Yanxia Sun, Performance analysis of cost-sensitive learning methods with application to imbalanced medical data, Informatics in Medicine Unlocked, Volume 25, 2021,100690, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2021.100690.

[7] Wan X, Liu J, Cheung WK, Tong T. Learning to improve medical decision making from imbalanced data without a priori cost. BMC Med Inf Decis Making Dec. 2014; 14(1):111. https://doi.org/10.1186/s12911-014-0111-9.

[8] Gan D, Shen J, An B, Xu M, Liu N. Integrating TANBN with cost sensitive classification algorithm for imbalanced data in medical diagnosis. Comput Ind Eng Feb. 2020;140:106266. https://doi.org/10.1016/j.cie.2019.106266.

[9] *Wang H, Cui Z, Chen Y, Avidan M, Abdallah AB, Kronzer A. Predicting hospital readmission via cost-sensitive deep learning. IEEE ACM Trans Comput Biol Bioinf Dec. 2018;15(6):1968–78. https://doi.org/10.1109/TCBB.2018.2827029.*

[10] Wu J-C, Shen J, Xu M, Liu F-S. An evolutionary self-organizing cost-sensitive radial basis function neural network to deal with imbalanced data in medical diagnosis. Int J Comput Intell Syst. Oct. 2020;13(1):1608–18. [Online]. Available: https://doi.org/10.2991/ijcis.d.201012.005.

[11] Chawla, N. V., Japkowicz, K., Kotzba, S., & Wróbel, W. (2002). Editorial: Special issue on learning from imbalanced data sets. ACM SIGKDD Explorations Newsletter, 6(1), 1-6. (https://www.researchgate.net/

[12] Breiman,L.(2001). Random forests. Machine learning, 45(3), 5-32. (https://link.springer.com/article/10.1023/A:1010933404324)

[13] Sun, Y., Wong, A. K., & Kwok, J. T. (2009). Cost-sensitive learning for multi-target prediction. In Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 875-884). (https://jds.acm.org/vol_1_issue_2.html)

[14] Huang, H., Sun, G., Hussain, Z., & Zhang, C. (2016). Deep imbalanced learning for face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4547-4555). (https://ieeexplore.ieee.org/iel7/34/4359286/08708977.pdf)

[15] XGBoost Documentation: https://xgboost.readthedocs.io/ (Look for the section on "scale_pos_weight" parameter)

[16] Zadrozny, B., Langford, J., & Abe, N. (2002). Cost-sensitive learning by cost-proportionate example weighting. In Third International Conference on Data Mining (ICDM'02) (pp. 107-114). IEEE. (https://hunch.net/~jl/projects/reductions/costing/finalICDM2003.pdf)

[17] Similar to cost-sensitive decision trees, you can use the reference by Zadrozny et al. (2002) mentioned above.

[18] Liu, W., Wang, Y., Li, S., Ling, H., & Lin, W. (2019). Cost-sensitive deep learning for imbalanced image classification. Neurocomputing, 354, 107-118. (https://www.sciencedirect.com/science/article/pii/S0925231219304151)