Journal of Information Systems and Telecommunication
Vol.13, No.3, July-September 2025, 177-188
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Resolving Class Imbalance in Medical Classification: Technique Comparison and Performance Evaluation

Abdallah Maiti[1*], Mohamed Hanini[1], Abdallah Abarda[2]

[1].Laboratory of Computing, Networks, Mobility and Modelling (IR2M) FST, Hassan First University of Settat, Morocco
[2].Laboratory LM2CE, Faculty of Economic Sciences and Management, Hassan First University of Settat, Morocco

Received: 16 Mar 2025/ Revised: 07 Aug 2025/ Accepted: 06 Sept 2025

Abstract

The problem of unbalanced data is a common one in medical diagnostics. This problem can reduce the accuracy of classification models and affect the validity of results. The aim of our paper is to compare several techniques for correcting class imbalances in medical datasets and to evaluate the impact of these techniques on machine learning performance.

In our paper, we used an imbalanced dataset to train a convolutional neural network (CNN) model. We then tested correction techniques such as sampling and cost-sensitive learning. Finally, we used recall, precision, accuracy and F1 score to evaluate the model's performance.

The results show that the use of correction techniques led to a significant improvement in the performance of the classification model. The cost-sensitive learning technique gave the best results, particularly for the detection of minority classes. This method increased the weight of classification errors associated with minority classes, thus improving the detection of critical cases. The results of this study underline the importance of dealing with imbalances in the data to improve the performance of classification models in the medical field. The use of methods such as cost-sensitive learning not only improves model performance, but also enables more reliable decisions to be made, which is essential for ensuring more accurate diagnoses and better quality of care.

Keywords: Data Imbalance; Techniques for Resolving Data Class Imbalance; Oversampling; Cost-Sensitive learning, Convolutional Neural Networks; Classification; Model Performance; Medical Diagnostics.

1- Introduction

The text must be in English. Authors whose English The problem of imbalanced data represents a big challenge in machine learning, particularly in critical fields such as healthcare, finance, cybersecurity and other. It occurs when certain classes in a data-set are underrepresented relative to others, causing predictive models to disproportionately favor the majority classes. In domains such as fraud detection, where fraudulent transactions represent only a small proportion of the data, models often struggle to identify these minority instances, favoring normal transactions instead [1], [2]. Similarly, rare diseases in medical diagnosis or infrequent cyberattacks in cybersecurity are often misclassified due to their limited representation in training datasets [3]. Addressing this imbalance is essential to improve prediction accuracy and ensure fairness across all classes. Classical ML algorithms, such as logistic regression and decision trees assume a balanced distribution of data, a condition that is rarely met in real-world applications. Therefore, various methods have been developed to mitigate biases caused by imbalance.

Different techniques such as oversampling, undersampling, cost-sensitive learning, and ensemble methods have shown promise in improving minority class detection while maintaining overall model performance [4] solve this problem. Imbalance can take different forms depending on the data type. In binary classification, a single minority class often poses a problem, as seen in rare disease diagnosis or fraud detection, where models tend to favor the majority class. Approaches such as SMOTE address this problem by generating synthetic examples for underrepresented categories [5]. In multi-class scenarios, imbalance arises when multiple classes are unequally represented, as seen in multi-stage disease diagnosis. In such cases, advanced techniques such as One-vs-One (OvO) and One-vs-Rest (OvR), as well as ensemble methods, are needed to ensure balanced performance across classes [4].

Beyond accuracy, traditional evaluation metrics often fail to capture a model's ability to identify minority classes.

✉ **Abdallah maiti**
abdallah.maiti@uhp.ac.ma

Metrics like precision, recall, and F1-score are more appropriate for binary imbalances, while G-mean and Matthews correlation coefficient (MCC) provide a more balanced evaluation for multi-class problems [6]. These metrics are crucial for evaluating mitigation strategies and ensuring fair representation of all classes.

Despite the progress made, significant challenges persist in combating class imbalance. Low performance on minority classes, inadequacy of conventional metrics, and difficulties in generalizing to unseen data are among the main obstacles. The choice of the most effective method depends on the specific context, including the severity of the imbalance and the area of application. In complex scenarios, hybrid approaches that combine data-level and algorithmic methods are often required [7].

Recent empirical investigations have underscored the efficacy of hybrid methodologies that integrate oversampling techniques, such as Synthetic Minority Over-sampling Technique (SMOTE), deep neural networks, and reinforcement learning to more proficiently address imbalance within intricate datasets. These adaptive methodologies are structured to correspond with the data's inherent architecture, thereby enhancing performance while concurrently mitigating the risk of overfitting [8]. Furthermore, the intensifying focus on algorithmic equity, especially within critical sectors like healthcare, necessitates the rectification of biases stemming from underrepresented classes, as such biases may precipitate significant diagnostic inaccuracies [8].

In the domain of natural language processing, contemporary scholarship regarding the Central Kurdish language has demonstrated that the qualitative balancing of corpora is imperative for guaranteeing the dependability of morphosyntactic frameworks, particularly in contexts characterized by limited resources [9].

These theoretical frameworks have significantly guided the methodological framework of the current investigation. The proposed architecture is predicated on a convolutional neural network (CNN), augmented by rebalancing methodologies such as Synthetic Minority Over-sampling Technique (SMOTE), classification paradigms including One-vs-One (OvO) and One-vs-Rest (OvR), alongside cost-sensitive learning and the ensemble-based Bagging methodology. This comprehensive framework aims to enhance the identification of minority classes while maintaining consistent overall efficacy.

In addition to extant research, this investigation enriches the academic discourse by amalgamating all four methodologies within a cohesive framework explicitly tailored for medical imaging applications. It delineates a multiclass classification protocol that tackles the infrequency of clinical cases, the hierarchical organization of disease stages, and the imperatives of algorithmic equity. This contribution is particularly notable in its deployment for the automated identification of diabetic retinopathy utilizing retinal imagery, where advanced stages of the condition are frequently underrepresented and challenging to discern.

The overall aim of this research is to develop a robust classification system capable of accurately identifying rare stages of diabetic retinopathy (DR). More specifically, the study seeks to determine the most effective techniques for correcting class imbalance in medical imaging; to evaluate the impact of these techniques using appropriate performance metrics such as recall and F1-score; and to offer practical recommendations for high-stakes domains where misclassification can significantly affect decision-making. The article is structured as follows: Section 2, “Materials and Methods,” describes the dataset, the CNN architecture, and the imbalance-handling strategies implemented; Section 3, “Results,” presents the model’s performance under various conditions; Section 4, “Discussion,” interprets the findings and considers methodological trade-offs; and finally, Section 5, “Conclusion,” summarizes the main contributions and proposes future research directions.

2- Materials and Methods

In our article, we investigate various techniques to address class imbalance in multi-class classification tasks. Our goal is to classify retinal images according to the severity stages of diabetic retinopathy (DR), a serious eye disease resulting from prolonged hyperglycemia. The dataset used is from the Kaggle platform and consists of five classes, ranging from “No DR” (absence of disease) to “Proliferative DR” (advanced and severe form of the disease). Unlike other studies that apply imbalance correction techniques without sufficient justification, we propose a systematic approach tailored to imbalanced and unstructured data, particularly images. Our aim is to scientifically identify the most effective techniques to overcome this challenge and evaluate their impact on the performance of classification models. To achieve this, we used a convolutional neural network (CNN)-based model, known for its ability to automatically extract complex features from images. We evaluate several class rebalancing techniques, including undersampling, oversampling, One-vs-Rest (OvR) and One-vs-One (OvO) approaches, cost-sensitive learning, and ensemble bagging (Fig.1). Models are trained and evaluated on balanced datasets using these techniques. The evaluation phase relies on standard metrics such as accuracy, precision, recall, and F1 score, which are derived from the confusion matrix. This comprehensive approach enables a precise analysis of the influence of the applied imbalance resolution techniques on the performance of the CNN-based model and provides insights into effectively addressing imbalances in image classification tasks.

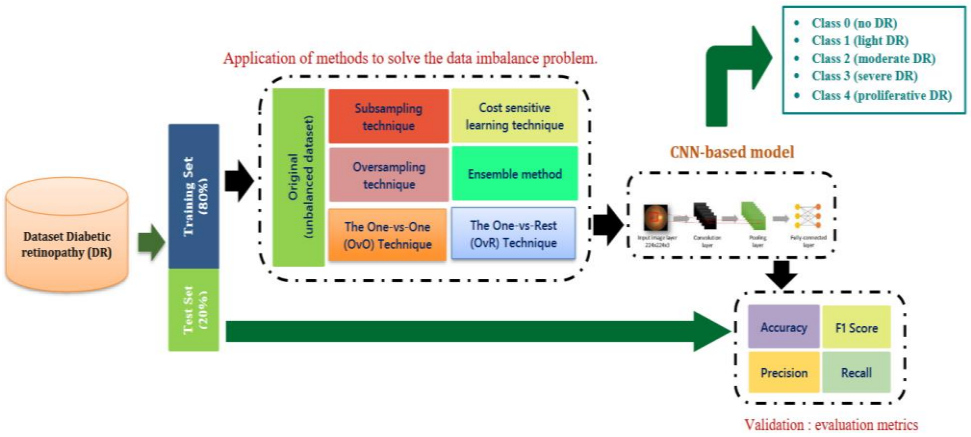

Fig. 1. Architecture of the proposed diagnostic system

2-1-Dataset Description

The dataset used in our paper and obtained from the Kaggle platform [29], consists of a total of 92702 retinal images distributed across five classes, each representing a stage of diabetic retinopathy (DR). The dataset (Table 1) exhibits a significant class imbalance, with the majority class, "No DR," comprising approximately 77.8% of the total samples. In contrast, the more severe stages, such as "Severe DR" and "Proliferative DR," are severely underrepresented, together accounting for less than 5.1% of the dataset.

Table 1. Distribution of Retinal Images Across Diabetic Retinopathy Classes

Class	***Description***	***Samples***	***Percentage***
Class 0	No DR	72102	77.8%
Class 1	Mild DR	8772	9.5%
Class 2	Moderate DR	7135	7.7%
Class 3	Severe DR	2328	2.5%
Class 4	Proliferative DR	2365	2.5%
Total		92702	100%

This imbalance poses challenges for model training, as predictive models tend to favor the majority class, leading to poor detection rates for minority classes. Addressing this issue is critical to improving diagnostic accuracy, particularly for the advanced stages of DR. Techniques such as oversampling, undersampling, and algorithmic adjustments are essential to mitigate this problem and ensure balanced and robust model performance.

2-2-Model Architecture

To solve the problem of multi-class classification of diabetic retinopathy, we have developed a model based on a convolutional neural network (CNN). This type of model is particularly effective for image analysis, thanks to its ability to automatically extract complex features while reducing the need for manual data pre-processing (Fig. 2).

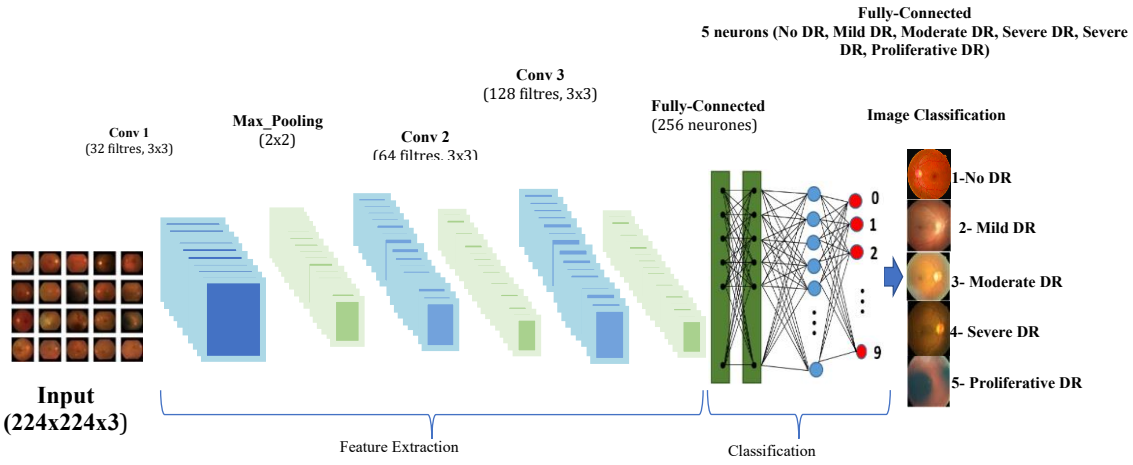

Fig. 2. Architecture of our CNN-based classification model

The architectural framework of the model is predicated upon a convolutional neural network (CNN) organized into three primary phases: feature extraction, dimensionality reduction, and classification. It consists of three convolutional layers designed to extract fundamental features from images, succeeded by pooling layers that facilitate dimensionality reduction and bolster the robustness of the model. Ultimately, two fully connected layers conclude the multi-class classification process. Methodologies such as dropout regularization, in conjunction with non-linear activation functions (ReLU and Softmax), augment the model's efficacy and generalizability in the identification of diabetic retinopathy.

2.2.1. Three Convolutional layers

The proposed model employs a triad of convolutional layers to derive critical features from retinal imagery. The initial layer utilizes 32 filters, succeeded by 64 filters in the subsequent layer and 128 filters in the final layer. Each filter executes a convolution operation utilizing a 3x3 kernel, thereby facilitating the identification of distinct patterns, including anomalies or textures that are characteristic of retinopathy.

2.2.2. Pooling layers (2x2)

After each convolutional layer, pooling layers with a 2x2 size kernel are applied to reduce the dimensionality of the data. This process limits over-fitting while reducing computational costs. The max-pooling method is used, selecting the maximum value in each analyzed region. This ensures that the most dominant and significant features of the images, essential for classification, are retained, while simplifying the representations learned by the model.

2.2.3. Two Fully Connected layers

The model comprises two fully-connected layers that ensure the finalization of the classification. The first layer, made up of 256 neurons, combines the features extracted from the convolutional and pooling layers. It uses a ReLU (Rectified Linear Unit) activation function, well known for its ability to introduce non-linearity, essential for modeling complex relationships between features. This function also prevents the effect of gradient saturation, which promotes efficient convergence during training.

The output layer comprises 5 neurons, corresponding to the five severity classes of diabetic retinopathy. A Softmax activation function is applied to transform the outputs of this layer into normalized probabilities, allowing direct interpretation of predictions as probabilities belonging to each class. This configuration is particularly well-suited to multi-class classification, guaranteeing well-calibrated output and a sum of probabilities equal to 1.

2.2.4. Regulation

A dropout mechanism (with a rate of 0.5) is implemented subsequent to the fully connected layers in order to mitigate the probability of overfitting by sporadically deactivating certain neurons throughout the training process. This methodology entails the random inactivation of 50% of the neurons at each iteration during training, thereby diminishing the model's excessive dependence on particular neurons.

This architecture integrates efficient convolutional layers for the automatic extraction of pertinent features alongside dense layers designated for classification. Such a framework is exceptionally well-suited for medical image analysis endeavors, owing to its capacity to capture intricate details while simultaneously minimizing the necessity for manual pre-processing.

2-3-Techniques for Correcting Data Imbalances

Addressing data imbalance is crucial for improving the performance of machine learning models. The different approaches to tackle this issue can be represented in three categories: data-driven approaches, algorithmic approaches, and specific approaches designed for multi-class problems.

2.3.1. Data-Based Methods

Data-based approaches involve the direct manipulation of datasets to balance the distribution of classes before model training.

a-Sub-Sampling

The technique of subsampling, unlike oversampling, involves reducing the number of samples from majority classes to balance their proportion relative to minority classes (Fig. 3). This technique is typically implemented by randomly removing examples from the dominant class [10]. Subsampling has several advantages, including model simplification by reducing the total volume of data, which also lowers computational costs. However, this technique has several notable drawbacks. Removing samples from majority classes can lead to the loss of crucial information [11]. Furthermore, the random selection of samples to be removed may not accurately reflect the actual distribution of the data, potentially affecting model performance, especially when the data is heavily unbalanced [12].

b-Oversampling

Oversampling methodologies pertain to the deliberate augmentation of sample quantities from minority classes to rectify their inadequate representation in imbalanced datasets (Fig 3). Among the preeminent methodologies, the Synthetic Minority Oversampling Technique (SMOTE) is particularly noteworthy for its capability to produce

synthetic instances through linear interpolation of existing samples within the minority class [6],[7]. This approach enhances the representation of underrepresented classes while concurrently maintaining the diversity and structural integrity of the dataset.

The practice of oversampling confers several advantages. It mitigates the model's bias towards majority classes and enhances its generalization capabilities. These benefits culminate in an improved recognition of underrepresented classes, particularly in scenarios where imbalances may precipitate erroneous predictions [13]. Furthermore, by infusing greater variability into minority classes, methodologies such as SMOTE enable machine learning algorithms to more effectively discern the unique characteristics of rare instances. Nonetheless, oversampling is not devoid of limitations. The artificial augmentation of samples may heighten the risk of overfitting, especially when synthetic instances exhibit insufficient diversity or replicate patterns that do not accurately reflect authentic data [14]. In addition, this escalation in data volume may incur elevated computational costs, particularly with extensive datasets, due to the supplementary resources necessitated for the generation and processing of synthetic instances [15]. Recent studies suggest improvements to SMOTE, such as K-Means SMOTE or Borderline-SMOTE, which specifically target critical regions near decision boundaries to maximize the efficiency of oversampling [16]. These variants aim to reduce drawbacks while fully exploiting the potential of minority classes in unbalanced contexts.

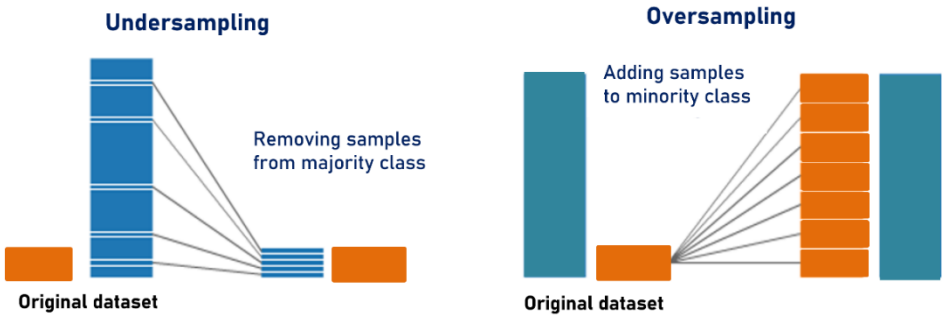

Fig. 3. Representative diagram of the two techniques: subsampling and oversampling

2.3.2. Algorithmic Approaches

Algorithmic approaches directly modify learning algorithms to deal with data imbalance, without modifying the distribution of classes in the ensemble.

a- Cost-Sensitive learning

This methodology modifies the loss function of machine learning algorithms by allocating enhanced significance to minority classes. This approach is predicated on augmenting the weight of errors pertinent to these classes, in accordance with their under-representation (Fig. 4). In a dataset wherein a class constitutes 10% of the samples, misclassification errors for that class may be amplified by a factor that corresponds to the degree of imbalance, thus escalating the associated penalty [17].

This methodology proves to be particularly efficacious in critical domains, such as the detection of rare diseases, the prevention of financial fraud, or the prediction of failures in intricate systems. It substantially contributes to the reduction of classification errors in under-represented classes, while simultaneously preserving the equilibrium of overall model performance [18]. In addition, by integrating these weights into algorithms, cost-sensitive learning augments model sensitivity and precision for imbalanced datasets.

Nonetheless, the efficacy of this methodology is profoundly contingent upon the meticulous calibration of the weights allocated to various classes. Insufficient calibration may result in an inverse imbalance, thereby impairing performance on majority classes or diminishing the overall effectiveness of the model [19]. Therefore, methodologies such as adaptive weight optimization or the employment of specific metrics, including the ROC curve or F-measure, are frequently advocated to guarantee balanced performance.

Fig. 4. Operating principle of the cost-sensitive learning method

b- Ensemble Methods

Ensemble techniques, such as Bagging and Boosting, combine the predictions of multiple models to enhance overall performance and reduce bias toward majority classes (Fig. 5). Bagging (Bootstrap Aggregating) uses random sampling with replacement to train several independent models, whose predictions are then aggregated, improving model robustness and stability [20]. Boosting, on the other hand, progressively corrects the errors of successive models by assigning higher weights to misclassified examples, thereby increasing overall accuracy, particularly on minority classes [21]. These techniques are particularly effective for datasets with a high degree of imbalance, as they address the weaknesses of individual models by improving the recognition of under-represented classes. By introducing diversity into data subsets and combining the strengths of several models, they also promote better generalization. Furthermore, recent variants, such as AdaBoost-SAMME or Gradient Boosting with SMOTE, have demonstrated their effectiveness in handling complex imbalances by adjusting weights for minority classes [23].

Nevertheless, the execution of these methodologies may prove to be intricate and computationally intensive, particularly in the context of boosting. The latter necessitates meticulous calibration of hyperparameters, including but not limited to learning rate and quantity of estimators, to mitigate the risk of overfitting and to guarantee optimal efficacy [24]. In spite of these obstacles, their capacity to enhance performance in scenarios characterized by imbalanced data renders them indispensable instruments in domains such as finance, healthcare, and predictive analytics.

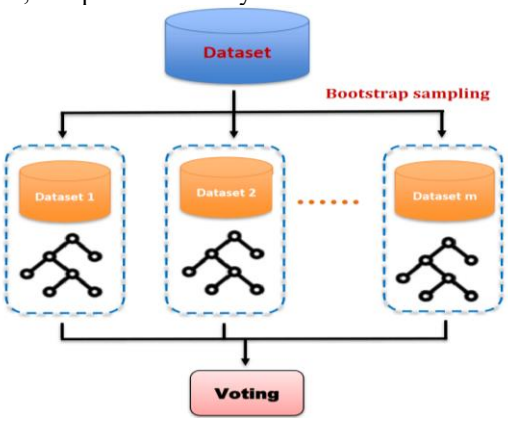

Fig. 5. Operating principle of the Bagging ensemble method

2.3.3. Specific Techniques for Multi-Class Problems

In multi-class problems, where multiple categories are present, data imbalance poses additional challenges. Classical approaches can be adapted, but specific approaches such as One-vs-Rest (OvR) and One-vs-One (OvO) (Fig. 6) are often used.

a- One-vs-Rest (OvR)

OvR also known as One-vs-All, decomposes a multi-class problem into several binary classification problems. For each class, a binary classifier is trained, treating this class as positive and grouping all other classes as negative. For instance, in a five-class problem, OvR requires the creation of five binary models, each optimized to distinguish a specific class [25],[26]. Notable advantages of this technique include its simplicity of implementation and its ability to provide independent evaluations for each class. These features make it particularly suited to contexts where granular predictions are essential, such as in image recognition or recommender systems [25],[26]. Additionally, the OvR technique is compatible with a wide range of learning algorithms, such as support vector machines (SVMs) and logistic regression, making it a versatile option.

However, this technique has important limitations. It can become biased when classes grouped as negative are highly imbalanced, which can impair model performance on minority classes [27]. Furthermore, OvR does not account for the complex relationships and possible interdependencies between different classes, limiting its ability to capture global patterns or subtle correlations in the data [28].

Recent work proposes extensions to mitigate these limitations, such as integrating adaptive weights to balance negative classes or using hybrid techniques that combine OvR with dimensionality reduction methods like linear discriminant analysis. These improvements aim to enhance the robustness and accuracy of this technique in unbalanced multi-class classification contexts.

b- One-vs-One (OvO)

The OvO technique treats each pair of classes separately, creating a binary classifier for each combination of two classes. For example, for a problem with five classes, the OvO results in ten binary classifiers, one for each pair of classes [25],[26].

This approach is particularly useful for data with complex class relationships, as each classifier focuses on only two classes at a time. This reduces the impact of majority classes, as each binary classifier works on data balanced between the two classes concerned. However, the computational complexity is high. The number of classifiers to be trained increases quadratically with the number of classes, which can lead to considerable computational costs and implementation difficulties in contexts with a large number of categories [27].

Data imbalance correction methods offer a variety of solutions tailored to specific application needs. Data-driven techniques, such as oversampling and undersampling, directly modify the class distribution, while algorithmic approaches, such as cost-sensitive learning and ensemble methods, adjust the algorithms to compensate for biases [28]. In multi-class problems, specific techniques such as

OvR and OvO are used to handle the additional complexity associated with multiple classes. The choice of the optimal method depends on the context of use, the nature of the data and technical constraints. It is often advisable to combine several approaches to maximize model performance while minimizing imbalance bias [25],[26].

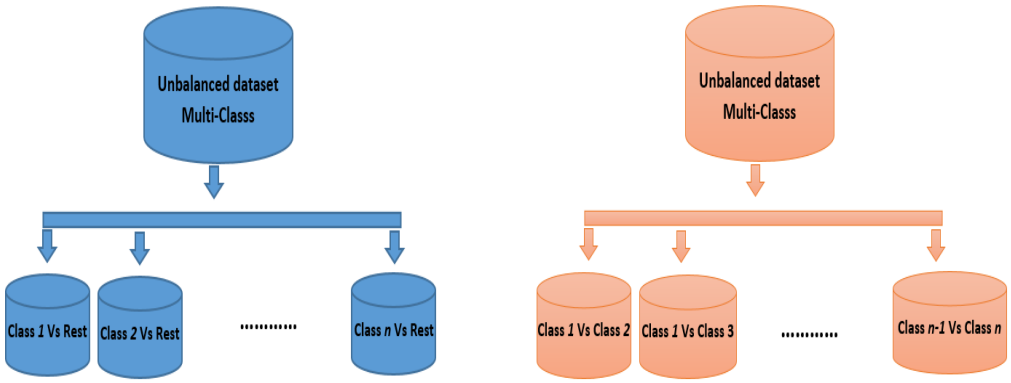

Fig. 6. Representation of the "One-vs-Rest"(OvR) and "One-vs-One"(OvO) techniques

3- The Results

Unbalanced multi-class classification is a major challenge, due to the complexity of interactions between classes and the difficulty of assessing model performance. Unlike binary classification, this context requires advanced approaches to effectively manage imbalance while improving prediction accuracy.

In our research, we apply and evaluate various data rebalancing techniques, such as oversampling, undersampling, one-to-one and one-to-all approaches, ensemble methods such as Bagging, and cost-sensitive learning. The aim is to identify the best method for boost the performance of artificial intelligence models in this complex context.

3-1-Subsampling

Sub-sampling is a methodological approach aimed at equilibrating the distribution of classes by diminishing the magnitude of the majority class, which is accomplished through the stochastic elimination of samples from this class to render it congruent with the quantity of the minority class. In the present investigation, each class was systematically curtailed to 2328 samples, in alignment with the size of the minority class. While this methodology serves to mitigate the bias in favor of the majority class, it engenders a considerable loss of information, which may adversely influence the overall efficacy of the model, as delineated in Table 2.

The implementation in Python employs the resample function from the sklearn.utils library to perform subsampling on the majority class, thereby modifying its size to correspond with that of the minority class. Subsequent to the subsampling procedure, the equilibrated dataset is preserved in the variables X_resampled and y_resampled, rendering it suitable for utilization in model training. The outcomes of this methodology are illustrated in Table 2.

Table 2. Overall performance obtained using the sub-sampling technique

Metric	*Global values*
Accuracy	82.64 %
Precision	88.94 %
Recall	82.15 %
F1-Score	80.51 %

3-2-Oversampling

To improve the representation of minority classes in unbalanced datasets, the SMOTE (Synthetic Minority Oversampling Technique) technique was used. SMOTE generates synthetic samples for under-represented classes by creating intermediate points between existing instances of the same class [30],[22]. This rebalances the distribution of classes and mitigates biases linked to data imbalance when training machine learning models.

In Python, SMOTE is implemented using the SMOTE class in the imbalanced-learn library (imblearn).

The resulting oversampling led to a significant improvement in overall performance, although there remains a risk of model overfitting due to the generation of synthetic samples. The performance results obtained after applying SMOTE are presented in Table 3.

Table 3. Overall performance obtained using the oversampling technique

Metric	*Global values*
Accuracy	87.09 %
Precision	84.36 %
Recall	81.78 %
F1-Score	83.05 %

The F1-Score of 83.05%, which combines two parameters: precision and recall into a single metric, provides a more comprehensive evaluation in handling imbalanced data. Although the accuracy is relatively high at 87.09%, it is not the most reliable metric for this type of task due to the potential influence of class imbalance. The moderate recall and F1-Score suggest that, while oversampling improved class distribution, the model may exhibit overfitting, limiting its ability to generalize effectively to unseen data.

3-3-Cost-Sensitive learning

Cost-sensitive learning is an effective technique for managing class imbalance without directly modifying the data distribution. It assigns weights proportional to the inverse of class frequency, thus giving greater importance to minority classes during training. In this study, weights were calculated as in Table 4.

Table 4. Weight of diabetic retinopathy classes

Class	*Weight*
Class 0	1
Class 1	(72 102/8 772) ≈ 8.22
Class 2	(72 102/7 135) ≈ 10.10
Class 3	(72 102/2 328) ≈ 31.00
Class 4	(72 102/2 365) ≈ 30.49

The weights were integrated into the SparseCategoricalCrossentropy loss function of TensorFlow/Keras through the class_weight parameter, thereby facilitating the equilibrium of performance between predominant and subordinate classes. This methodology dynamically modifies the error magnitude associated with under-represented classes, obviating the necessity for direct alterations to the training dataset, and empowers the model to more effectively manage class imbalances during the training process.

In this specific implementation, the class_weight parameter is employed to modulate the significance of each class, thereby compensating for imbalances while preserving the integrity of the data itself. Metrics such as Accuracy, Precision, Recall, and F1-Score were computed on the test dataset to appraise the model's efficacy. Upon the completion of training the CNN-based model, its performance was evaluated utilizing the test data (refer to Table 5). The findings illustrate that this methodology proficiently reconciles overall accuracy and performance across all classes, including minority classifications, thereby mitigating the adverse effects of data imbalance on predictive quality. The model accomplished an Overall Accuracy of 91.09%, indicative of its capacity to render precise predictions across all classifications. The F1-Score, a composite metric amalgamating precision and recall, attained 92.79% for the "No DR" classification, underscoring the model's dependability in identifying this category. Below is a comprehensive delineation of the performance metrics for each class:

No DR: The model exhibited outstanding performance in this category, attaining a Precision of 91.14%, a Recall of 94.49%, and an F1-Score of 92.79%, which exemplifies its robust capability to accurately recognize instances devoid of diabetic retinopathy. Mild DR: This classification similarly exhibited elevated performance, achieving a Precision of 93.27%, a Recall of 91.95%, and an F1-Score of 92.60%, signifying a well-balanced aptitude for detecting mild cases. Moderate DR: With a Precision of 91.95%, a Recall of 93.24%, and an F1-Score of 92.59%, the model effectively identified moderate cases with negligible errors. Severe DR: The performance of the model was somewhat diminished for this classification, achieving a Precision of 88.26%, a Recall of 82.86%, and an F1-Score of 85.47%, which reflects certain challenges in differentiating severe cases. Proliferative DR: This minority classification attained a Precision of 85.88%, a Recall of 83.72%, and an F1-Score of 84.78%, demonstrating the model's capacity to address even the most formidable cases, albeit with some constraints.

Table 5. Performance obtained by applying Cost Sensitive Learning

Metric	*Overall Accuracy*	*Precision*	*Recall*	*F1-Score*
No RD	91.09 %	91.14 %	94.49 %	92.79 %
light RD		93.27 %	91.95 %	92.60 %
Moderate RD		91.95 %	93.24 %	92.59 %
Severe RD		88.26 %	82.86 %	85.47 %
Proliferative RD		85.88 %	83.72 %	84.78 %

3-4-Ensemble technique: Bagging

Bagging (Bootstrap Aggregating) was implemented in Python to handle unbalanced data sets. Four balanced subsets were created by bootstrap sampling, each subset comprising 2,328 representative samples of all classes, including minority classes, using scikit-learn's resample function. These subsets were used to independently train a CNN model, developed with TensorFlow using a defined architecture, an 'adam' optimizer, a 'categorical_crossentropy' loss function, and 'accuracy' metrics.

The predictions of the four models were aggregated by majority voting, implemented via scipy's mode function. The results obtained are presented in Table 6.

Table 6. Overall performance of the Bagging technique

Metric	*Global values*
Accuracy	83.21 %
Precision	83.49 %
Recall	83.21 %
F1-Score	83.28 %

3-5-OvR and OvO Techniques

OvR and OvO techniques are widely used strategies for handling multi-class classification problems, particularly when addressing class imbalance. In this study, these techniques were implemented in Python.

The overall performance of these two techniques is summarized in Table 7.

Table 7. Overall performance achieved using OvR and OvO techniques

Technique	*Accuracy*	*Precision*	*Recall*	*F1-Score*
OvR	84.06 %	80.35 %	83.53 %	81.91 %
OvO	79.68 %	81.65 %	84.19 %	82.90 %

The results show that the OvR technique achieves an accuracy of 84.06%, while OvO performs better in terms of precision and F1-Score, albeit with slightly lower accuracy. These two techniques are complementary, and the choice of approach will depend on the specific objectives of the model, notably between precision and recall.

4- Discussion

Table 8. presents the performance of the CNN classification model, trained on the “DR” (Diabetic Retinopathy) dataset balanced by different techniques. This table compares the results obtained with different class imbalance correction techniques, assessing their impact on four main metrics: Accuracy, Precision, Recall and F1-Score.

This comparison highlights the strengths and limitations of each technique, as well as their influence on overall model performance.

The comparative results of the different imbalance correction techniques are shown in Table 8. above. The metrics used (Accuracy, Precision, Recall and F1-Score) make it possible to evaluate the effectiveness of each technique on overall model performance.

a- Cost-Sensitive Learning Technique

The cost-sensitive learning methodology modifies the weightings assigned to each class in accordance with their prevalence, thereby effectively mitigating biases resulting from class imbalance. Among the methodologies assessed, cost-sensitive learning demonstrates the most favorable overall efficacy, yielding an accuracy of 91.09%, a precision of 90.10%, a recall of 89.25%, and an F1-score of 89.65%. This approach is particularly adept at addressing the disparate costs associated with misclassification, enabling the model to more accurately identify minority classes while preserving elevated overall precision. The exemplary outcomes of cost-sensitive learning illustrate its capacity to reconcile precision and recall, rendering this technique an outstanding selection for datasets characterized by imbalance. While the performance metrics are commendable, it is crucial to acknowledge that the dynamic recalibration of weights may incur significant computational costs, particularly when engaging with extensive datasets. Our findings regarding cost-sensitive learning align with those reported in contemporary scholarly literature, which has evidenced that this strategy stands out as one of the most efficacious for imbalanced multi-class classification challenges, as evidenced by the research conducted by Khan et al. [31]. A more recent investigation by Araf et al. [32] posits that this technique necessitates meticulous parameter optimization to circumvent computational burdens while sustaining high precision. This highlights the imperative for practitioners to diligently evaluate the trade-offs between computational expenses and performance enhancements.

b- Oversampling Technique

Oversampling, particularly using the SMOTE method, generates synthetic samples for minority classes, improving their representation during training. SMOTE achieved an

accuracy of 87.09%, precision of 84.36%, recall of 81.78%, and an F1-score of 83.05%. While this method is powerful, it carries the risk of overfitting if the synthetic data does not accurately reflect the complexity of real samples.

It is important to note that the risk of overfitting can be a major issue with this approach. According to Vargas et al. [33], the generated samples may introduce unrealistic variations into the data, which could harm the model's ability to generalize. This trade-off between improving the representation of minority classes and the risk of overfitting must be carefully evaluated.

c- Bagging Technique

Bagging (Bootstrap Aggregating) significantly bolsters the reliability of predictions through the amalgamation of numerous models that have been trained on meticulously balanced subsets of the dataset. This methodology attained an accuracy rate of 87.49%, a precision level of 84.91%, a recall metric of 81.72%, and an F1-score of 83.28%. While it exhibits a marginal advantage over oversampling with respect to accuracy, the computational resources required for training multiple models may pose a limitation in environments constrained by resources. Despite the robustness of this technique, the substantial computational demands must be meticulously evaluated. As posited by Liang & Zhang [34], the process of training various models on data subsets necessitates effective resource management, which can serve as an impediment in computationally limited scenarios. Consequently, the balance between precision and computational expense must be critically assessed in professional practice.

d- Subsampling Technique

Under-sampling entails the reduction of the population of the majority class to correspond with the population size of the minority classes. This methodology yielded an accuracy rate of 82.64%, a precision rate of 88.94%, a recall rate of 82.15%, and an F1-score of 85.41%. Although this methodology facilitates the equilibrium between precision and recall, it is plagued by a considerable diminution of information, which may adversely influence the model's capacity to generalize.

The information attrition linked to under-sampling can detrimentally affect the generalization capabilities of the model, as articulated by Soleimani & Mirshahzadeh [35]. In real-world implementations, this strategy may prove to be suboptimal when substantial amounts of information are essential for the accurate prediction of infrequent occurrences, as is the case with diabetic retinopathy.

e- OvO and OvR Methods:

The One-vs-One (OvO) and One-vs-Rest (OvR) methodologies partition the multi-class classification challenge into binary subproblems. The efficacy of the OvO method is marginally inferior to that of alternative methodologies, attaining an accuracy of 79.68%, a precision of 81.65%, a recall of 84.19%, and an F1-score of 82.90%. Conversely, the OvR methodology achieves an accuracy of 84.06%, yet it remains suboptimal in performance relative to strategies such as cost-sensitive learning and oversampling. Our findings regarding OvR and OvO are in alignment with those documented in contemporary research, including the work of Chakraborty & Dey [36], which indicates that while these methodologies may be effective in certain contexts, they are generally less efficacious than approaches like cost-sensitive learning (CSL) and Synthetic Minority Over-sampling Technique (SMOTE) due to the inherent trade-offs in accuracy and computational efficiency.

Table 8. Model performance on the balanced DR dataset using different imbalance correction techniques

Correction techniques	***Accuracy***	***Precision***	***Recall***	***F1-Score***
Subsampling	82.64 %	88.94 %	82.15 %	85.41 %
Oversampling	87.09 %	84.36 %	81.78 %	83.05 %
Cost-sensitive learning	**91,09%**	**90,10%**	**89,25%**	**89,65%**
Bagging technique	87.49 %	84.91 %	81.72 %	83.28 %
One-vs-One (OvO)	79.68 %	81.65 %	84.19 %	82.90 %
One-vs-Rest (OvR)	84.06 %	80.35 %	83.53 %	81.91 %

5- Conclusion

The categorization of images depicting diabetic retinopathy poses a considerable challenge attributable to class imbalance, a widespread concern within medical applications. This manuscript conducts a comparative analysis of diverse methodologies aimed at mitigating this imbalance while simultaneously enhancing the efficacy of Convolutional Neural Network (CNN) models. The findings unequivocally indicate that the selection of correction methodologies exerts a substantial influence on model efficacy, thereby underscoring the necessity for the adoption of strategies that are specifically tailored to the contextual characteristics of the data and the distinct aims of the application.

Among the methodologies scrutinized, cost-sensitive learning emerges as the preeminent strategy. Its adaptive modulation of class weights facilitates a balanced evaluation of classification inaccuracies, culminating in enhanced performance across critical metrics (Accuracy, Precision, Recall, and F1-Score). This approach not only assures superior generalization but also yields a more precise identification of minority classes. Techniques such as oversampling and bagging also exhibited favorable outcomes, particularly in augmenting the representation of minority classes, while concurrently sustaining competitive overall performance. Nonetheless, both methodologies may engender a compromise between computational expense and precision, particularly in expansive applications. Conversely, subsampling and the One-vs-One/One-vs-Rest (OvO/OvR) techniques, although beneficial, are encumbered by intrinsic limitations, such as potential information loss or heightened complexity, rendering them less appropriate for intricate, imbalanced datasets such as those associated with diabetic retinopathy.

These observations accentuate the imperative for a comprehensive evaluation of the strengths and weaknesses inherent to each technique, with particular emphasis on the trade-offs between computational expenditure and accuracy. The outcomes further highlight the significance of implementing solutions specifically adapted to the particular constraints of the data and the objectives of the application. Future investigations should prioritize the innovation of novel methodologies that effectively manage complex, imbalanced datasets. Additionally, the exploration of hybrid models that amalgamate existing techniques should be pursued to capitalize on the synergistic strengths of each strategy. This integrative methodology would contribute to the optimization of performance by addressing the deficiencies associated with individual techniques, thereby enhancing model capabilities in regard to both accuracy and generalization.

Such a strategy would not only elevate the overall performance of models but also more effectively address the critical requirements of applications, particularly in domains such as medicine, where the robustness, fairness, and reliability of models are of paramount importance.

References

[1]. KrawczykB, B. (2016). “Learning from imbalanced data: Open challenges and future directions”. Published in Progress in Artificial Intelligence, V5(4), pp 221-232.

[2]. Haixiang, G., and al. (2017). “Learning from class-imbalanced data: Review of methods and applications”. Published in Expert Systems with Applications, v73, pp 220-239.

[3]. LemaîtreG., Nogueira, F., and Aridas, C. K(2017). « Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning”. Published in Journal of Machine Learning Research, v18(17), pp1-5.

[4]. BrancoP., Torgo, L., andRibeiro, R. P2019). A survey of predictive modeling on imbalanced domains. ACM Computing Surveys, v49(2), pp1-50.

[5]. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.

[6]. ChawlaN. V. et al2002). SMOTE: Synthetic Minority Over-sampling Technique. Published in Journal of Artificial Intelligence Research, 16, 321-357.

[7]. Kaur, H. et al. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. Published in ACM computing surveys (CSUR), 52(4), 1-36.

[8]. Abdullah, A. A., Mohammed, N. S., Khanzadi, M., Asaad, S. M., Abdul, Z. K., & Maghdid, H. S. (2025). In-depth Analysis on Machine Learning Approaches: Techniques, Applications, and Trends. ARO-THE SCIENTIFIC JOURNAL OF KOYA UNIVERSITY, 13(1), 190-202.

[9]. Sabr, S. S., Mustafa, N. S., Omar, T. S., Rasool, S. H., Omer, N. A., Hamad, D. S., ... & Maghdid, H. S. (2025). A Comprehensive Part-of-Speech Tagging to Standardize Central-Kurdish Language: A Research Guide for Kurdish Natural Language Processing Tasks. arXiv preprint arXiv:2504.19645.

[10]. Kaur, H., Pannu, H. S., and Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. ACM computing surveys (CSUR), 52(4), 1-36.

[11]. LinC. C., Yen, S. J., and Lee, Y. S2017). On combining SMOTE with under-sampling: An experimental study on class imbalance problem. Published in Information Sciences, v371, 123-137.

[12]. YangC., at al. (2024). Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. Published in Journal of big data, v11(1), 7.

[13]. Loffredo, E., Pastore, M., Cocco, S., & Monasson, R. (2024). Restoring balance: principled under/oversampling of data for optimal classification. arXiv preprint arXiv:2405.09535.

[14]. Buda , M. , Maki, A., and Mazurowski, M. A. (2018). “A systematic study of the class imbalance problem in convolutional neural networks”. Neural Networks, 106, pp 249-259.

✉ **Abdallah maiti**
abdallah.maiti@uhp.ac.ma

[15]. Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. Journal of Big Data, 5(1), 1-30.
[16]. Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. Advances in Intelligent Computing, 878-887.
[17]. Liu, Y., Wu, T., & Yan, P. (2020). Balancing imbalanced data using adaptive synthetic sampling with feature selection. Computational Intelligence and Neuroscience, 2020, 1-11.
[18]. Longadge, R., & Dongre, S. (2013). Class imbalance problem in data mining review. arXiv preprint arXiv:1305.1707.
[19]. Brownlee, J. (2020). Imbalanced classification with Python: better metrics, balance skewed classes, cost-sensitive learning. Machine Learning Mastery.
[20]. Yadav, S., & Bhole, G. P. (2020, December). Handling imbalanced dataset classification in machine learning. In 2020 IEEE Pune Section International Conference (PuneCon) (pp. 38-43). IEEE.
[21]. Liu, L., Wu, X., Li, S., Li, Y., Tan, S., & Bai, Y. (2022). Solving the class imbalance problem using ensemble algorithm: application of screening for aortic dissection. BMC Medical Informatics and Decision Making, 22(1), 82.
[22]. Maiti, A., Abarda, A., & Hanini, M. (2022, October). A New Hybrid Artificial Intelligence Model for Diseases Identification. In The Proceedings of the International Conference on Smart City Applications (pp. 825-836). Cham: Springer International Publishing.
[23]. He, H., Garcia, E. A. (2009). "Learning from imbalanced data". In IEEE Transactions on knowledge and data engineering, 21(9), pp1263-1284.
[24]. Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification : an experimental review. Journal of Big data, 7, 1-47.
[25]. Pawara, P., Okafor, E., Groefsema, M., He, S., Schomaker, L. R., & Wiering, M. A. (2020). One-vs-One classification for deep neural networks. Pattern Recognition, 108, 107528.
[26]. Brownlee, J. (2020). One-vs-rest and one-vs-one for multi-class classification. Machine Learning Mastery.
[27]. LiQ., SongY., ZhangJ., and ShengV. S2020). « Multiclass imbalanced learning with one-versus-one decomposition and spectral clustering". Published in Expert Systems with Applications, in147, p113--152.
[28]. Chakraborty, S., & Dey, L. (2024). Multi-class Classification. In Multi-objective, Multi-class and Multi-label Data Classification with Class Imbalance: Theory and Practices (pp. 51-76). Singapore : Springer Nature Singapore.
[29]. Diabetic Retinopathy Detection data set, in kaggle.com/c/diabetic-retinopathy-detection/data
[30]. Maiti , A., Abarda, A., Hanini, M., and Oussous, A. (2024). "An Optimal Model Combining SqueezeNet and Machine Learning Methods for Lung Disease Diagnosis. Current Medical Imaging, 20(1).
[31]. Khan, A. A., Chaudhari, O., & Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. Expert Systems with Applications, 244, 122778.DOI : 10.1016/j.eswa.2023.122778
[32]. Araf, I., Idri, A., & Chairi, I. (2024). Cost-sensitive learning for imbalanced medical data: A review. Artificial Intelligence Review, 57(4), 80.DOI : 10.1007/s10462-023-10652-8
[33]. Vargas, W. de, Schneider Aranda, J. A., dos Santos Costa, R., da Silva Pereira, P. R., & Victória Barbosa, J. L. (2023). Imbalanced data preprocessing techniques for machine learning: A systematic mapping study. Knowledge and Information Systems, 65(1), 31-57.DOI : 10.1007/s10115-022-01772-8
[34]. Liang, G., & Zhang, C. (2012). A Comparative Study of Sampling Methods and Algorithms for Imbalanced Time Series Classification. In AI 2012: Advances in Artificial Intelligence (pp. 637–648). Springer.DOI : 10.1007/978-3-642-35101-3_54
[35]. Soleimani, M., & Mirshahzadeh, A. S. (2023). Multi-class classification of imbalanced intelligent data using deep neural network. EAI Endorsed Transactions on AI and Robotics, 2, 1-10.DOI : 10.4108/airo.7998.
[36]. Chakraborty, S., & Dey, L. (2024). Applications of Multi-objective, Multi-label, and Multi-class Classifications. In Multi-objective, Multi-class and Multi-label Data Classification with Class Imbalance: Theory and Practices (pp. 135-164). Singapore: Springer Nature Singapore. DOI : 10.1007/978-981-97-9622-9.