Addressing Class Imbalance in Healthcare Data: Machine Learning Solutions for Age-Related Macular Degeneration and Preeclampsia

Antonieta Martínez-Velasco, Lourdes Martínez-Villaseñor, and Luis Miralles-Pechuán

***Abstract*—The use of machine learning in healthcare has transformed the way diseases are diagnosed and treatments are optimized. However, medical databases often lack balanced data due to challenges in data collection caused by privacy regulations. Certain health conditions are underrepresented, which hampers machine learning performance. To address this problem, a hybrid approach has been proposed that combines the Synthetic Minority Oversampling Technique (SMOTE) with undersampling and uses two specific techniques tailored for imbalanced datasets. Comparative evaluations were conducted using various thresholds to reduce one class and employing Balanced Accuracy to mitigate bias toward the majority class, with popular machine learning methods. The results showed that Balanced Bagging and Balanced Random Forest consistently outperformed other methods, performing the best with an average ranking of 1.42 and 3.58 out of 32 configurations in the two datasets, respectively. Tree-based approaches such as Random Forest and Gradient Boosting demonstrated similar effectiveness, emphasizing the power of aggregating predictions from multiple trees to reduce bias. Notably, undersampling and SMOTE proved advantageous for non-tree-based models like KNN, SVM, and Logistic Regression, showcasing their usefulness across different algorithms. This study provides a robust solution for handling imbalanced datasets in healthcare, which could potentially optimize healthcare interventions and improve patient outcomes and care.**

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/8952

***Index Terms*—Machine Learning Techniques, Healthcare Domain, Class Imbalance, Ensemble Classifiers, Diagnostic Decision-Making, Personalized Medicine.**

I. Introduction

Machine Learning techniques are methods and approaches that enable computers to learn from data and make predictions or decisions. These techniques analyze information, generate predictions, and automate various tasks.

Over recent decades, Machine Learning (ML) has made remarkable strides, finding successful applications in diverse domains [1]. ML plays a pivotal role in facilitating the development of automated decision-making systems by uncovering patterns within expansive datasets. In this context, diagnostic decision-making is aided by ML techniques to identify the underlying cause of a patient's symptoms or condition. It involves gathering information, analyzing data, and applying medical knowledge to reach a diagnosis. ML has brought revolutionary changes in disease diagnosis, treatment, and prediction [2].

ML has significantly impacted the Healthcare Domain, a vast field encompassing health services, industries, and practices. Its applications span disease diagnosis, treatment, prevention, education, wellness programs, and health system management.

Dealing with the healthcare sector presents distinctive challenges, especially in the intricate process of acquiring patient data. Obtaining data from healthy individuals adds an extra layer of complexity, requiring a more exhaustive search effort than the relatively accessible patient data. The lower engagement of healthy individuals with healthcare systems further complicates assembling a representative control group.

As a result, medical datasets often exhibit a significantly higher number of cases than control groups, as elucidated in a study by khushi et al. [3]. The complexity of gathering data, especially for conditions like cancer, demands strict adherence to rigorous ethical protocols sanctioned by relevant authorities and hospitals. These protocols are crucial to ensuring data privacy and confidentiality protection, as emphasized by the guidelines provided by Centro del Conocimiento Bioetico in 2015 [4].

The paradigm of personalized medicine, emphasizing prevention, early diagnosis, and targeted treatment, demands reliable screening, diagnosis, and prognosis methods to identify at-risk patients and initiate appropriate interventions [5]. Medical records, often complex and incomplete data repositories, are the primary source for identifying disease risk factors. Unfortunately, these databases frequently exhibit imbalanced class distributions, hindering effective knowledge extraction [5].

Class imbalance arises in ML datasets when one class significantly outnumbers the other. This uneven distribution can hinder the training of accurate models, particularly for the underrepresented classes [6]. The prevalence of imbalanced class distributions in medical studies yields unstructured, multi-modal data with potential biases [7], [8]. This imbalance issue can significantly impact the performance of standard classifiers designed under the assumption of balanced datasets, thereby leading to biased outcomes favoring the majority class

The associate editor coordinating the review of this manuscript and approving it for publication was Suélia Fleury (Corresponding author: Antonieta Martínez-Velasco).

A. Martínez-Velasco, and L. Martínez-Villaseñor are with the Facultad de Ingeniería, Universidad Panamericana, Augusto Rodin 498, Ciudad de México, 03920, México (e-mails: amartinez, and lmartine @up.edu.mx).

L. Miralles-Pechuán is with Faculty of Computer Science, Technological University Dublin, Grangegorman, Dublin, Ireland (e-mail: luis.miralles@tudublin.ie).

[9]. The potential consequences are severe, compromising the accuracy of algorithms and posing risks to patient well-being if misclassifications occur [10]. Exploring alternative solutions becomes imperative in the face of challenges associated with increasing dataset sizes.

This study addresses the challenge of unbalanced healthcare data by implementing data-level and algorithmic techniques. Focusing on two complex diseases, Age-Related Macular Degeneration (AMD) and Preeclampsia, the research investigates how the proposed methods can support expert decision-making. The comprehensive process involves preprocessing data to handle missing values, selecting relevant features, and utilizing resampling techniques to address class imbalance.

This paper presents key contributions in addressing this issue, focusing on a novel hybrid sampling technique, case studies on disease datasets, exploration of configuration impact, a comprehensive comparison of classification methods, and efficiency analysis. The main contributions of the paper are:

- **Hybrid Sampling Technique:** This paper introduces a novel hybrid approach that integrates oversampling and undersampling in supervised ML classifiers for addressing class imbalance in medical databases. By combining random oversampling and the Synthetic Minority Oversampling Technique (SMOTE), the proposed method enhances the algorithm performance in healthcare datasets.
- **Case Studies on Disease Datasets:** The study conducts comprehensive case studies on Age-Related Macular Degeneration and Preeclampsia datasets, showcasing the effectiveness of the proposed approach. Specifically, the robust performances of Balanced Bagging and Balanced Random Forest, exceeding the other methods' balanced accuracy in Age-Related Macular Degeneration cases, demonstrate the practical application of the hybrid technique in real-world healthcare scenarios.
- **Exploration of Configuration Impact:** The research explores six distinct case reduction configurations (0%, 50%, 60%, 70%, 80%, and 90%) with and without both SMOTE and Undersampling, providing valuable insights into the impact of specific classification methods and sampling techniques on handling class imbalance in healthcare datasets. This analysis contributes to a deeper understanding of optimal configurations for different scenarios.
- **Comparison of Classification Methods:** The study systematically compares various classification methods with different sampling techniques in addressing class imbalance. Notably, Balanced Bagging consistently outperforms other methods, showcasing its reliability, while the adaptability of Gradient Boosting and Decision Tree methods underscores their effectiveness in diverse healthcare scenarios.
- **Efficiency and Runtime Analysis:** The research evaluates the efficiency and runtime of the proposed methods, revealing that Balanced Bagging, despite its computational cost, remains the most effective method with Balanced Random Forest performing slightly worse but with a notably shorter runtime. This efficiency analysis contributes practical considerations for implementing the proposed approach in real-world healthcare applications.

The subsequent sections of this article are organized as follows: Section II provides an overview of the two case studies, AMD and Preeclampsia. It also reviews current techniques addressing the problem of unbalanced classes, categorized into data-level and algorithm-level techniques. Section III outlines the proposal to address class imbalance in both case studies and describes the dataset and methodology. Section IV presents the obtained results from these methods. Section V analyses the obtained results of the experiments and highlights some important observations. Finally, Section VI concludes the article, emphasizing the significance of exploring techniques for imbalanced datasets in the health sector.

II. State of the Art

This section provides an overview of ML applications in AMD and Preeclampsia. Below is a concise exploration of how ML techniques have been employed in medical research to address complex issues related to the prognosis and diagnosis of both diseases. Finally, this section explores the challenges posed by imbalanced datasets in machine learning classification. It discusses the limitations of conventional classification techniques in handling class imbalance and introduces two main approaches to address this issue: Data-Level Solutions and Algorithm-Level Solutions.

A. ML Approaches to Study Age-Related Macular Degeneration

AMD is a major cause of vision loss in older individuals, targeting the central retina, which is crucial for sharp vision, as shown in Fig. 1. By 2020, AMD was projected to affect approximately 196 million people worldwide, with numbers potentially increasing to 288 million by 2040. This condition can cause significant visual impairment, interfering with daily tasks such as reading and driving. There are two forms of AMD: dry, which is more prevalent and progresses gradually, and wet, which is rarer but leads to quicker and more severe vision loss [11].

AMD is a prevalent cause of visual impairment and blindness, affecting both developed and underdeveloped countries. In the United States, approximately 9% of individuals aged over 65 have AMD, rising to 28% in those over 75 [12]. In Mexico, 1,241,000 people experienced vision loss in 2016, equivalent to 1.01% of the population [13]. AMD, along with glaucoma, stands as a major cause of irreversible vision loss. AMD primarily affects the macula, resulting in gradual central vision loss, often characterized by the accumulation of drusen, deposits in the macula leading to late-stage indicators like geographic atrophy or sub-retinal neovascularization [14].

There are two types of risk models for AMD: prediction and inference. Prediction models aim to provide the best possible risk assessment by combining genetic, non-genetic, and clinical factors. These models are not widely accepted in the medical community, and the American Academy of Ophthalmology (AAO) has stated that genetic testing for complex diseases like AMD will not be routine until clinical trials can

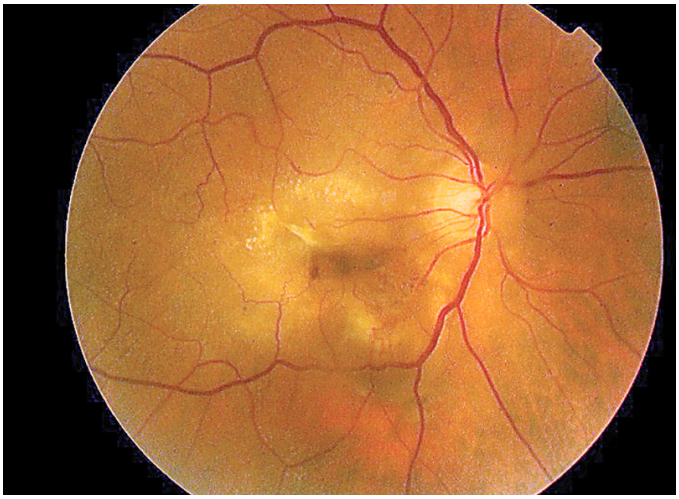

Fig. 1. A 75-year-old man experienced a gradual distortion of vision in his right eye, which eventually progressed to a loss of central vision. Source:Flickr.com

demonstrate that patients with specific genotypes benefit from particular types of therapy or surveillance [15]. The complexity of AMD, including the different clinical phenotypes and other confounding factors, makes it challenging to develop accurate predictive tests. Classical statistical models have been the primary method for modeling AMD. However, studies have recently emerged focusing on risk factors, particularly genetic variants, due to advances in genome scanning [16].

ML techniques identify intricate patterns in a given dataset, allowing for prediction and inference in new datasets without human intervention. In the clinical vision sciences, these techniques are commonly used to study retinal diseases [17]. Two approaches have been used by scientists to study retinal diseases. One involves studying retinal images to associate them with risk factors. The other approach uses ML for explainability, which aims to provide medical professionals with more information on how the model works. For instance, ML has been used to extract knowledge on identifying malignant mesothelioma by providing rules that explain which factors are more relevant for diagnosing it [18].

Studies that explore the association between genetic and environmental risk factors aim to predict diseases. Scientists attempt to identify the genes and polymorphisms involved in complex diseases since the hereditary component is essential. In genome-wide association studies (GWAS), scientists select a gene considered a risk factor and then analyze a few of its polymorphisms to determine the association between its alleles and a phenotype or the frequency of a disease [19]. Every year, GWAS are published with increasing associations of single nucleotide polymorphisms (SNPs) with diseases or phenotypes [16].

Further statistical analysis is necessary to find a polymorphism or variant of a gene associated with a disease in a specific population. Firstly, scientists study familial aggregation to determine if the disease is genetically determined. Secondly, they must locate genes of interest for the disease. In the identified areas, there may be thousands of polymorphisms of interest [20].

Traditionally, probability and statistical techniques have been used to understand complex diseases. The difficulty of handling large amounts of data due to its volume, speed, and variability makes current methods computationally unfeasible. However, ML techniques can aid scientists in finding predictors that are related to the onset of the disease [21]. Personalized medicine aims to pre-symptomatically identify people at high risk of disease using knowledge of the person's genetic profile and their environmental risk factors [22]. Therefore, forecasting made through ML can represent a support tool for medical professionals [23].

AMD has been studied from various angles. While the study of images obtained through ocular fundus studies has been widely explored, a less explored approach considers genetic variants in addition to other risk factors. In Martínez-Velasco et al. [24], a review of the studies conducted to find the risk factors associated with AMD with both approaches has been presented.

ML is mentioned as an important tool to transform the enormous volume of complex data into knowledge under the approach of personalized medicine. Larrañaga et al. [25] present some of the most useful techniques for bioinformatics modeling and optimization and highlight the application of ML methods in genetics. Spencer et al. [26] propose to increase the datasets by using the multi-factorial dimensionality reduction (MDR) and the grammatical evolution of neural networks (GENN), in addition to the logistic regression (LR) approach. Combining the LR and GENN model results, the algorithm achieves a sensitivity of 77.0% and specificity of 74.1%. Jiang et al. [27] propose a random forest (RF) adaptation for epistatic interactions. The sliding window epi-Forest algorithm incorporates RF in case-control studies and automates screening candidate polymorphisms for statistical analysis to detect epistatic interactions.

ML approaches are presented as complementary methods to facilitate the exploration of interactions between multiple polymorphisms because epistasis plays an important role in the pathogenesis of AMD. The authors proposed the importance of Gini, obtained from AMD classification methods, which can complement the p-value of statistical studies to measure associations between polymorphisms and AMD and to identify possible combinations of genetic variants that are protective for AMD. Gold et al. [28] analyzed a set of risk factors for AMD with a statistical model and then with an MA model. This was done with a model based on Genetic Algorithms (GA).

The advantage of GAs over traditional models is their ability to incorporate multiple positions across the genome to make a prediction. This allows models to identify complex interactions between polymorphisms correlated with their predictions. Chen et al. [29] proposed a method based on sets of trees to identify the interactions between genes and the interactions between genes and the environment. This method is proposed to solve the problem of missing data and for the selection of attributes simultaneously. This approach avoids the problem of collinearity for genome-wide data and does not require any a priori assumption.

Forest-based algorithms are a popular ML tool due to their adaptability to data, applicability to big and small problems, and ability to consider correlations and interactions between entities. One major concern in identifying disease genes is the occurrence of false positives. The authors of a study [29] were able to successfully distinguish regions of the genome associated with the disease from neutral regions with a false positive rate (FPR) of less than 5%. Çelebiler [30] explored the relationship between the presence of multiple genetic polymorphisms, risk factors, and dry and wet AMD.

Three types of Bayesian networks were constructed to investigate the relationship between the presence of multiple genetic polymorphisms and AMD. Bayesian networks help with the learning process by incorporating prior knowledge and cause less noise and over-fitting than other methods. The flexibility of the models allows for precise decisions to be made from uncertain data. Fraccaro et al. [31] compared "white box" models (such as logistic regression and decision trees), which are most interpretable, with "black box" methods such as support vector machine (SVM), Random Forest, and Adaptive boost.

Both methods, white and black boxes, identified soft drusen and age as the most important variables in diagnosing AMD. The authors [31] emphasized the importance of interpretation and limiting the number of samples required to obtain reliable results for early diagnosis. They proposed a graphical user interface to show the diagnostic pathway or variable importance, providing specialists with decision pathways to make early diagnoses feasible and better differentiate ambiguous subsets of patients. Krishnaiah et al. [32] presented the predictive performance of the artificial neural networks model in comparison with the predictive capacity of the logistic regression model.

B. ML Applied to Preeclampsia

The second scenario for applying ML techniques for imbalanced datasets in our study concerns Preeclampsia, a disease with significant implications for the global population. Preeclampsia is a severe pregnancy complication marked by high blood pressure and damage to organs, particularly the liver and kidneys. It impacts about 5-10% of pregnancies globally and is a major contributor to maternal and fetal illness and death. Typically developing after 20 weeks of pregnancy, preeclampsia can result in life-threatening complications if not managed. As shown in Fig. 2, the condition can lead to preterm birth, low birth weight, and increase the long-term risk of cardiovascular diseases for both mother and child.

The dataset used is interesting due to the high degree of imbalance. This data was obtained from the Inter-university Consortium for Political and Social Research [33] and aimed to improve children's health and well-being from birth to age three in Trenton, New Jersey. The program had three main strategies: (1) improving access to prenatal care and effective parenting, (2) improving the quality of child care, and (3) strengthening and maintaining positive parental involvement in their children's lives. Preeclampsia is defined as the onset of hypertension and proteinuria during the second half of pregnancy [34]. Clinical circumstances such as diabetes mellitus,

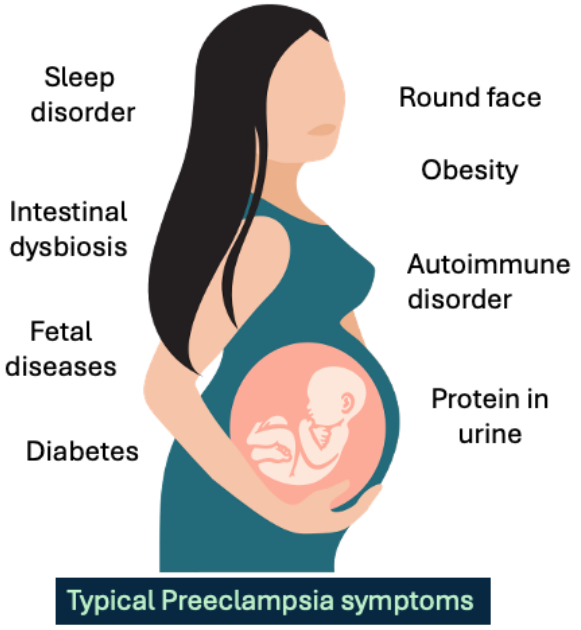

Fig. 2. Expecting mother facing the challenges of preeclampsia such as experiencing sleep disorders, elevated blood pressure, and concerns over fetal health and gestational diabetes.

obesity, systemic lupus erythematosus, and maternal age over 40 years are known to increase the risk of preeclampsia [35].

Since preeclampsia affects a large number of women worldwide, identifying demographic factors, biochemical analysis, or biophysical findings to predict the later development of preeclampsia in early pregnancy is crucial [36]. The high prevalence of preeclampsia is a problem that requires consideration, and predicting each case is essential. Therefore, a methodology must be devised to maximize the number of predicted cases adequately [37].

Mechanistic studies have provided information on the pathogenesis of the disease. They have also created opportunities to study circulating and urinary biomarkers to predict the disease [38]. An ML approach has been applied in various approaches, including metabolites, image analyses, and risk factors datasets, to diagnose and predict the disease. Kenny et al. [39] proposed a tree-based genetic programming method to diagnose PE by analyzing three metabolites in blood plasma. They presented demographic data and analyzed the concentration of some metabolites in both groups, cases, and controls. They showed results in a chromatogram, along with rules to obtain conclusions.

In a study, Neocleous [40] utilized Neural Networks to classify a database of 15 risk factors, achieving the best results with a multi-slab neural structure that could estimate the risk of PE occurring at an early stage. They graphed the neural networks' performance. Espinilla [41] classified a risk factors data set using decision trees without pruning and linguistic fuzzy transformation, using a genetic approach. They presented some rules generated by the decision tree. Velikova [42] proposed classification through the Bayesian Networks

model manually built using expert knowledge.

To provide relevant input data, a mobile app offers risk factors and measurements of signs. The author presents user interfaces showing the risk of getting the disease based on the temporal Bayesian network. Tejera et al. [43] classified a clinical history data set, including risk factors, to characterize PE. Results are presented regarding ROC curves, sensitivity, specificity variations, and Normalized Importance of the independent variables in the Artificial Neural Networks. Villa et al. [44] proposed Bayesian clustering to compute the risk ratio of each disease outcome.

This paper displays the results in the heat map, presenting the risk factors in different clusters. The system's decision-making process is not explicit. Moreira [45] proposed classifying risk factors, physiological mechanisms, and symptoms data set to identify high-risk pregnancies. This work proposed a smart system designed to support medical decision-making for pregnancy. Fergus [46] classified a genetic variants data set based on Genomic Wide Association Study (GWAS) using Deep learning stacked auto-encoders to allow early detection of preeclampsia. Additionally, they proposed using structured logic rules to reduce the interpretability of neural networking models. Cox [47] used Bayes Networks as the better algorithm to classify a plasma membrane proteins data set. Mehta reviewed and analyzed data mining methods applied to maternal care [48]. The author concluded that graphical representations of Decision Trees and Naïve Bayes models are easier to understand for medical experts, unlike Neural Networks and Vector Machines.

C. Unbalanced Datasets

Conventional classification techniques naturally adjust to the majority class, since most samples are over-represented in the loss function. Traditional regularization techniques, designed to balance bias and variance, do not regularize one-sided biases, such as over-fitting one class to the detriment of the other [49]. The problem of datasets with unbalanced classes has been approached from both the data level and algorithm level [50].

1) Data-Level Solutions: Data-level solutions are independent of the classification method [51]. Among these are resampling methods that focus on the data. These methods are categorized into sub-sampling, oversampling, and hybrid resampling. Table II provides a brief review of relevant works that offer data-level solutions.

Some methods involve resampling the data when dealing with class imbalance in datasets. One way to do this is by either sampling the minority or majority class until both have roughly equal representatives. However, both methods have their drawbacks. Subsampling can generate useful data, but oversampling can artificially increase the size of the dataset and create a heavier computational load. Both methods also alter the original distribution of the classes. A simple way to increase the size of the minority class is through random sampling, but this can result in over-fitting because it creates exact copies of minority class examples. Another strategy is to use a technique to generate new minority synthetic examples based on several positive examples that are very close together. Based on this review, a hybrid resampling method could be a potential solution to address class imbalance issues in healthcare datasets.

2) Algorithm-Level Solutions: The second approach to solving the class imbalance problem is an algorithmic-level solution to modify classification techniques to improve classification performance by adjusting the weights for each class. Algorithm-level solutions include sets of classifiers and cost-sensitive learning algorithms [73]. Table I presents a review of the most representative works to give an overview of the solutions at the algorithm level.

In conclusion, some algorithmic solutions, such as cost-sensitive approaches, depend more on the specific problem under study, while the data-level solutions and the classifier ensembles are more versatile [62]. This work proved the performance of a hybrid data-level solution for the balance in the health domain. The sub-sampling and oversampling techniques and ensembles were applied in two study cases: Age-Related Macular Degeneration and Preeclampsia.

III. Methodology

This subsection presents the pipeline for carrying out the experiments for the aforementioned techniques to two scenarios in the medical domain: Age-Related Macular Degeneration and Preeclampsia.

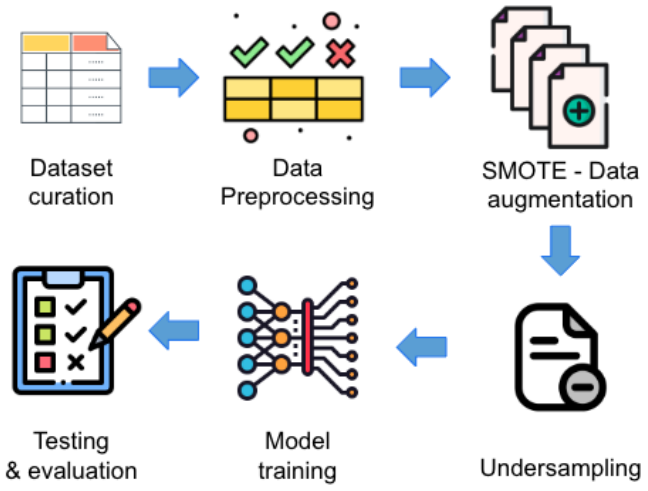

Fig. 3. Methodology Overview: A brief visual representation of the study's step-by-step process involving dataset preprocessing, feature selection, addressing imbalances, noise elimination, classifier set usage, and balanced accuracy evaluation.

The following procedure was followed:

1) Preprocessing of each data set to address the missing values problem.
2) Replacing missing values using the average and mode.
3) Apply RFE to select the optimal number of variables to build the models.
4) Applying different thresholds to remove instances from the class. The study applies thresholds from 50% to 90% at 10% intervals to the case class and when Undersampling is applied it does so to the control class.

TABLE I
ALGORITHM-LEVEL SOLUTIONS USED FOR BALANCING IMBALANCED DATASETS, INCLUDING AUTHORS, SPECIFIC METHODS EMPLOYED FOR TASK BALANCING, AND THE TYPES OF DATA EVALUATED IN DIFFERENT STUDIES

Author	Balancing Method	Data Type	Evaluation Metrics
Zhu et al. [52]	R library (IRIC) - a new implementation	Three datasets from the telecommunication industry	ROC
Wang et al. [53]	Classification: BSMAIRS	Eight medical public datasets	Accuracy, Sensitivity, Specificity, G-mean
X. Liu and Z. Zhou [54]	C4.5 decision tree, B-C45CS	Thirty-eight public datasets	Total cost
Khan et al. [55]	Cost-sensitive deep neural network (CoSen)	Six main image classification datasets	F-measure, G-mean
Weiss et al. [56]	"Budget-sensitive" progressive sampling algorithm	Twenty-six public datasets	Area under the ROC curve
McCarthy et al. [57]	Cost-sensitive learning and sampling	Twelve public datasets and two from AT&T	ROC
Japkowicz et al. [58]	C5.0, Neural Networks, Support Vector Machines	Eight public datasets	Classification percent error
Alanis-Tamez [59]	Assisted Classification for Imbalanced Data (ACID)	Fifteen public datasets of common chronic diseases	Friedman's test, Holm's test
Moreno [60]	Parallel classification system based on mixed-type assembly of experts (PCEM)	Ten public datasets	Accuracy
Xia et al. [61]	Optimization algorithm based on accelerated proximal gradient and block coordinate descent techniques	Public and real Cardiovascular and Cerebrovascular Disease datasets	Accuracy, Hamming Loss, Ranking Loss, F1
Diez Pastor et al. [62]	Various ensemble-based methods with diversity-increasing techniques	Eighty-four public datasets	AUC, F-Measure, G-Mean
Mena et al. [63]	Rule Extraction for Medical Diagnosis (REMED)	Four public datasets	Accuracy, Sensitivity, Specificity, AUC, G-Mean
Kumar et al. [14]	SMOTE, ADASYN, SVM-SMOTE, SMOTEEN, and SMOTETOMEK	Five public data sets	Accuracy, precision, recall
Li et al. [64]	Neural Network (DA-RNN) and Convolutional Block Attention Module (CBAM)	Public data set	Accuracy, root mean square error (RMSE), and the mean absolute error (MAE)
Santos et al. [65] 6	Artificial Immune Systems (AIS) and Decision Trees (DT) induced via Genetic Programming (GP)	Public data set	Sensitivity, specificity
Peng et al. [66]	Active learning and semi-supervised learning methods	Three synthetic, and one public data set	
Ullah et al. [67]	Deep learning method with a pre-trained deep residual network, ResNet-18	MIT-BIH arrhythmia database, INCART 12-lead Arrhythmia Database	Accuracy, precision, recall, F1-score
Mazur-Milecka et al. [68]	XGBoost, Support Vector Machine (SVM), Random Forest, and Explainable Boosting Machines (EBM)	Synthetic data set	Accuracy
Kovacheva et al. [69]	Xgboost, and linear regression	Data set from Mass General Brigham hospitals electronic health records (EHR)	AUC
Chłopowiec et al. [70]	Convolutional Networks	22 publicly available datasets	Sensitivity, specificity, accuracy, AUC, F1-score
Xie et al. [71]	Class imbalanced semisupervised learning (CISSL)	Synthetic dataset	Precision, recall
Veturi et al. [72]	Convolutional neural networks	Moorfields Eye Hospital (MEH) dataset	(AUROC) and Cohen's Kappa

5) Addressing the disequilibrium by applying resampling techniques.
6) Elimination of instances of the majority class.
7) Oversampling the minority class using SMOTE.
8) Use of sets of classifiers to improve the performance of individual classifiers.
9) Evaluation of the ML models using the balanced Accuracy as the performance metric.

The methods employed for the experiments are as follows: Balance Bagging [85], Balanced RF [86], Random Forests [87], Gradient Boosting [88], Decision Trees [89], Logistic Regression [90], K-nearest neighbor [91], and Support Vector Machine [92]. The Ensemble classifiers were applied to combine the predictions of multiple individual models to improve overall performance. This combination often leads to greater accuracy, robustness, and generalization than a single model.

Deep learning methods excel with images when there is an important number of instances. For example, they have been applied to unbalanced datasets, e.g., for detecting skin cancer [82]. However, they were not considered in this study due to their suboptimal performance in scenarios with small datasets, a common characteristic in the health sector where accessing records is challenging due to stringent patient data protection measures.

IV. EXPERIMENTS & RESULTS

We conducted experiments to assess how well the model performs when dealing with varying degrees of class imbalance. Our main evaluation metric was balanced accuracy, which measures the average accuracy across both positive and

TABLE II
DATA LEVEL SOLUTIONS USED FOR BALANCING AND CLASSIFYING IMBALANCED DATASETS, INCLUDING AUTHORS, BALANCING AND CLASSIFICATION METHODS, AND DATA TYPES

Author	Balancing & Classification Methods	Data Type	Evaluation Metrics
Batista [51]	Two methods: Smote + Tomek and Smote + ENN	Thirteen two-class public datasets with different degrees of imbalance	ROC curve (AUC)
Chawla [74]	SMOTE + combined sub-sampling method. Classifiers: C4.5, Ripper, and Naive Bayes	Nine datasets with varied sizes and class proportions, featuring continuous and nominal features	AUC, ROC convex hull strategy
Palodeto et al. [75]	Comparison of two methods: random over-sampling and random under-sampling. Classifier: artificial neural networks	Protein structure databases for Protein Secondary Structure Prediction (PSSP)	Evaluated prediction accuracy for each class of protein secondary structure and general accuracy
Blagus & Lusa [76]	SMOTE in high-dimensional data; KNN classifier	Simulated and real gene expression datasets	Predictive accuracy, AUC, G-mean
Rodríguez Torres [77]	Comparison of SMOTE-D against SMOTE and other over-sampling methods	Public datasets	AUC, F-Measure (F-M)
Vluymans [78]	Fuzzy rough set-based method for multi-class imbalanced classification (FROVOCO)	Eighteen multi-class imbalanced public datasets	Average accuracy, Multi-AUC
Mohammed et al. [79]	Three normalization methods (Min-Max, Z-Score, L2) along with SMOTE	Dataset with 7000 diabetic patients	Accuracy, Precision, Recall, F1, ROC curve
Yao et al. [80]	Data enhancement with rotation and mirroring,Swin Transformer	1200 posterior pole of the eye images	Accuracy, sensitivity, specificity, and F1-score
Khan et al. [81]	Undersampling,VGG-19	ODIR (Ocular Disease Intelligent Recognition) fundus images	Accuracy, precision, recall, and F1-score
Alam et al. [82]	Data augmentation techniques: rotation, and flipping.	Skin Cancer MNIST: HAM10000 dataset (Multisource dermatoscopic images)	Accuracy and ROC curve
Sowjanya et al. [83]	SMOTE modifications: Distance-based SMOTE (D-SMOTE) and Bi-phasic SMOTE (BP-SMOTE)	Framingham dataset from Wisconsin Hospital, and Novel Coronavirus 2019 dataset relating to forecasting COVID-19 cases	Accuracy
Koc et al. [84]	Random under-sampling (RUS), random over-sampling (ROS) and synthetic minority over-sampling technique (SMOTE)	Dataset collected from Turkey Social Security Institution	NA

negative classes, regardless of their distribution. To handle class imbalance, we used the SMOTE for both oversampling and undersampling.

Imbalance was systematically reduced from 50% to 90% in 10% intervals to demonstrate the robustness of our approach. To further test the methods against class imbalance, we leveraged specialized methodologies designed for skewed datasets. These included Balance Bagging [85] and Balanced Random Forest (Balanced RF) [86], known for their effectiveness in handling imbalanced data distributions. We chose not to use deep learning methodologies due to the relatively small size of our datasets. Deep learning has been recognized to show suboptimal performance on smaller and imbalanced datasets, which aligns with our experimental constraints and objectives [93].

A. Age-Related Macular Degeneration Case

The first step was to address any missing data to handle the data set properly. In the case of the AMD data set, missing data examples were replaced with either the mean or the median, depending on whether the variable was categorical or numeric, respectively. The number of missing values was only 1%, which did not result in any significant loss of information.

To select the most relevant variables for classifying the AMD data set, the Recursive Feature Elimination (RFE) algorithm was applied using Random Forest as the evaluator. Each variable was classified according to its importance to the model. The generated models were evaluated using the Accuracy metric to determine the predictive power of the set of variables in the classification process.

The RFE algorithm adjusted the model to the 29 variables in the data set. RFE tested all possible combinations and stored their performance in a variable combination list. For each iteration, all variables were reclassified. At the end of the algorithm execution, a sorted list was obtained using the results of all iterations. From the combination of five variables (Ophthalmology Surgeries, SNP rs203687, Bilateral Cataracts, Alcohol Intake, and SNP rs11200638 (Table III)). Therefore, experiments were conducted using all the features. The accuracy across all features can be seen in Fig. 4.

TABLE III
VARIABLE IMPORTANCE FOR AMD DATA SET

Variables	Accuracy	Kappa
Bilateral Cataract	0.8336	0.6718
CFH87 (rs203687 SNP)	0.8303	0.6659
Alcohol intake	0.8291	0.6622
HTRA1 (rs11200638 SNP)	0.8267	0.6567
Ophthalmology surgery	0.7971	0.6030

The AMD data set had a moderate imbalance, with an IR (Imbalance Ratio) of 0.8686, and the ratio between cases and controls is 0.4648 for cases and 0.5351 for controls.

TABLE IV
PERFORMANCE METRICS FOR AMD DATASET USING DIFFERENT CLASSIFICATION METHODS WITH VARIOUS SAMPLING TECHNIQUES. THE TABLE PRESENTS BALANCE ACCURACY SCORES FOR EACH METHOD UNDER DIFFERENT CONDITIONS

NUM	METHOD	SMOTE	Under Sampling	0%	50%	60%	70%	80%	90%	TIME (SEC)
1	Balanced Bagging	F	F	**0.9722**	**0.9722**	**0.9722**	**0.9722**	**0.9718**	**0.9175**	**1.50**
		F	T	0.9722	0.9722	0.9722	0.9722	0.9444	0.8833	3.50
		T	F	0.9722	0.9722	0.9722	0.9599	0.9425	0.831	7.17
		T	T	0.9722	0.9722	0.9722	0.9722	0.9438	0.8897	3.50
2	Balanced Random Forest	F	F	**0.969**	**0.9722**	**0.9722**	**0.969**	**0.9718**	**0.9128**	**4.33**
		F	T	0.9698	0.9722	0.9722	0.9705	0.9444	0.876	6.17
		T	F	0.9706	0.9718	0.9683	0.9405	0.9298	0.8077	14.17
		T	T	0.9698	0.9722	0.9718	0.9677	0.935	0.8866	9.00
3	Decision Tree	F	F	0.9556	0.9298	0.9258	0.925	0.9135	0.8512	18.83
		F	T	**0.9532**	**0.9528**	**0.9532**	**0.946**	**0.9373**	**0.9008**	**15.33**
		T	F	0.9536	0.9317	0.9313	0.9274	0.9103	0.877	18.50
		T	T	0.9532	0.9536	0.952	0.9524	0.9286	0.8976	15.83
4	Gradient Boosting	F	F	**0.9627**	**0.9774**	**0.9758**	**0.9623**	**0.9623**	**0.9623**	**5.17**
		F	T	0.9623	0.9698	0.9683	0.9623	0.9421	0.9008	11.17
		T	F	0.9635	0.9708	0.9634	0.9636	0.9345	0.8798	13.17
		T	T	0.9643	0.9679	0.9671	0.9495	0.9421	0.9017	11.67
5	KNN	F	F	**0.9583**	0.9008	0.8313	0.8214	0.7143	0.6786	25.17
		F	T	0.9583	0.8929	0.8671	0.8075	0.7917	0.8254	23.00
		T	F	**0.9567**	**0.9243**	**0.9128**	**0.9156**	**0.8839**	**0.8444**	**20.33**
		T	T	0.9538	0.907	0.8603	0.819	0.7915	0.8148	23.83
6	Logistic Regression	F	F	0.9107	0.8651	0.8294	0.8075	0.7381	0.6667	28.33
		F	T	0.9107	0.8829	0.869	0.7917	0.8948	0.7619	25.33
		T	F	0.9095	0.8571	0.8521	0.8359	0.8181	0.7172	27.00
		T	T	**0.911**	**0.9015**	**0.8743**	**0.8608**	**0.8687**	**0.7559**	**24.00**
7	Random Forest	F	F	**0.9702**	**0.9702**	**0.9651**	**0.923**	**0.9012**	**0.6845**	**17.33**
		F	T	0.9702	0.9722	0.9722	0.9704	0.9442	0.887	5.33
		T	F	0.971	0.9722	0.9694	0.9484	0.9402	0.8215	10.67
		T	T	0.9706	0.9722	0.9722	0.9687	0.9402	0.8776	6.83
8	SVM	F	F	0.8889	0.6548	0.5119	0.5	0.5	0.5	31.83
		F	T	0.8889	0.877	0.6567	0.5972	0.5258	0.5258	30.00
		T	F	**0.916**	**0.8858**	**0.8665**	**0.8642**	**0.8933**	**0.7807**	**24.00**
		T	T	0.923	0.7647	0.6263	0.5628	0.5263	0.5258	29.67

B. Experiments for the Preeclampsia Case

The dataset contains 1,640 records, and 12.4% of them have missing data. In this process, the missing data was replaced with the median or average value, depending on whether the variable was categorical or numerical. The first step was to identify the most important variables for predicting Preeclampsia and check for missing values in those variables. The mean or median replaces the missing values in these variables. After preprocessing the database, RFE with Random Forest was used to select the most relevant variables based on their importance in the model.

The accuracy of the generated models was evaluated to determine the predictive power of the selected variables in the classification process. The RFE algorithm tested all possible combinations and stored them in a variable combination list along with their performance. All variables were reclassified for each iteration to select the most relevant ones. To overcome the class imbalance problem, the receiver operational characteristics (ROC) curve was applied by selecting the Leave One Out cross-validation method. The ROC curve helped us select the most relevant variables by limiting the maximum possible value along the axis.

The variables that make up the Preeclampsia data set are listed in Table V according to the Modified Dynamic Gini Index (MDGI), which measures the inequality among values of a frequency distribution [94]. The MDGI value of each variable is expressed in the range [0, 100]. The most important variables, according to the MDGI, are the duration of pregnancy completed in weeks (PRGLNGTH), poverty (POVERTY), workforce status, water retention/edema in pregnancy (SWLNANKL), education (years of schooling completed) (EDUCAT), workforce status (LABORFOR), school or highest school grade (HIEDUC), and the number of cigars smoked a day 6 months before the woman knew she was pregnant (PRIORSMK). Based on the ROC curve, using all 25 variables to continue with the subsequent steps for data set classification. As was done with the previous dataset, the RFE algorithm was applied to the dataset. The results are shown in Fig. 5. For doing the experiments all variables were selected except one.

The dataset was unbalanced, with 269 instances for the positive class and 1371 for the negative class. It has a total of 26 variables. The IR is 0.1962, indicating the ratio of cases to controls. The Preeclampsia data set has an IR of 0.1962, resolved by the SMOTE algorithm.

After addressing the class imbalance issue, the next step was eliminating any noisy data. The Preeclampsia dataset mostly consists of categorical variables, with only five being numeric. The "One hot" encoding algorithm was used to transform each category into a binary vector with a corresponding numeric

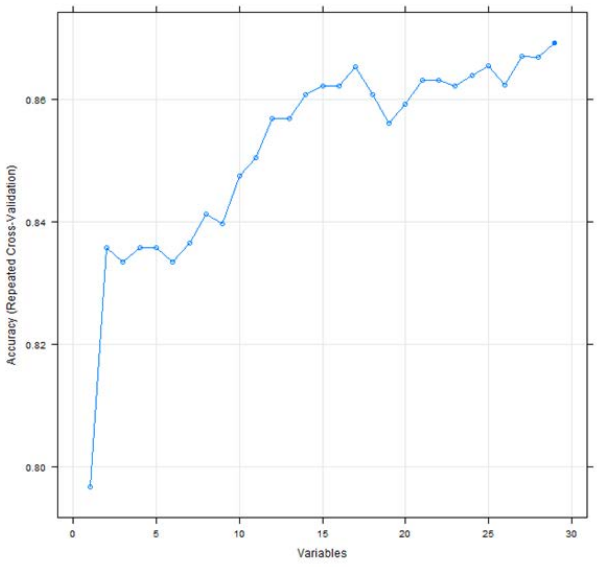

Fig. 4. Enhancing the classifier's accuracy through the RFE algorithm, using RF as an evaluator across various combinations.

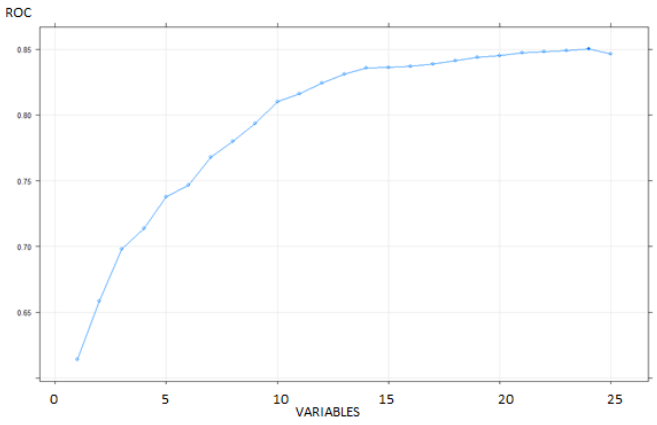

Fig. 5. Improving the accuracy of the classifier by employing the RFE algorithm, utilizing RF as an evaluator across different combinations.

value to prevent categorical variables from interfering with classification. The undersampling technique was implemented to randomly eliminate instances of the majority class, as done in previous experiments [95]. A total of 184 instances, which comprised 17.1% of the majority class, were deleted from the Preeclampsia dataset. With the problems of class imbalance, categorical variables, and spurious data resolved, the classification process could continue. The interclass imbalance in the Preeclampsia dataset was much larger than that of AMD.

V. Discussion of the Results

Experiments involved six distinct case reduction configurations (0%, 50%, 60%, 70%, 80%, and 90%), both with and without SMOTE and with and without Undersampling. This resulted in a total of 30 repetitions for each configuration.

In the analysis of the Age-Related Macular Degeneration (AMD) dataset, various classification methods with different

TABLE V
Preeclampsia Data Set Variables Importance According to the Modified Dynamic Gini Index (MDGI)

Variables	MDGI
PRGLNGTH	27.80
POVERTY	25.80
SWLNANKL	19.42
EDUCAT	11.14
LABORFOR	10.90
HIEDUC	10.36
PRIORSMK	9.40
OUTCOME	8.43
REGION	8.12
RELIGION	7.61
NEWPR	6.80
METRO	6.68
OTHRPROB	6.29
RACE	4.49
VGBLDFST	4.17
HISPRACE	4.10
ANEMIA	3.86
GESTDBTS	3.71
WEAKCRVX	3.11
NPOSTSMK	2.96
POSTSMKS	2.89
VGBLDLST	2.80

sampling techniques were employed to address class imbalance (Table IV). Balanced Bagging and Balanced Random Forest methods demonstrated robust performance across different thresholds, achieving balanced accuracy scores above 97% in some cases. Gradient Boosting exhibited competitive performance, consistently achieving scores above 96%. K-Nearest Neighbors (KNN) showed promising results, particularly when coupled with SMOTE, reaching up to 95.67%.

For the Preeclampsia dataset, similar analyses were conducted, focusing on balanced accuracy scores (Table **??**). Balanced Bagging and Balanced Random Forest again stood out, with balanced accuracy scores exceeding 77%. Decision Tree and Gradient Boosting methods demonstrated competitive results, particularly with the application of SMOTE, achieving balanced accuracy scores above 72%. Logistic Regression displayed notable performance, reaching a balanced accuracy of 74.12%, further enhanced by the use of SMOTE.

These findings underscore the effectiveness of specific classification methods and sampling techniques in handling class imbalance in healthcare datasets. The consistent performance of Balanced Bagging and Balanced RF highlights their reliability, while the adaptability of Gradient Boosting and Decision Tree methods emphasizes their effectiveness in diverse scenarios. LG, when paired with SMOTE, showcases its potential for optimizing model performance in imbalanced datasets.

A few interesting points about the experiments are:

- The most effective method is Balanced Bagging, albeit with a significant computational cost. Balanced Random Forest also performs admirably, albeit with a notably shorter runtime.

TABLE VI
PERFORMANCE METRICS FOR PRECLAMPSIA USING DIFFERENT CLASSIFICATION METHODS WITH VARIOUS SAMPLING TECHNIQUES. THE TABLE PRESENTS BALANCE ACCURACY SCORES FOR EACH METHOD UNDER DIFFERENT CONDITIONS

NUM	METHOD	SMOTE	Under Sampling	0%	50%	60%	70%	80%	90%	TIME (SEC)
1	Balanced Bagging	**F**	**F**	**0.7865**	**0.7788**	**0.782**	**0.7758**	**0.7736**	**0.7598**	**1.33**
		F	T	0.786	0.7597	0.7495	0.7538	0.7212	0.7002	6.50
		T	F	0.6872	0.7513	0.7623	0.7722	0.7721	0.7354	5.00
		T	T	0.6857	0.6574	0.6493	0.6643	0.6628	0.6455	15.83
2	Balanced Random Forest	**F**	**F**	**0.7714**	**0.7776**	**0.7693**	**0.7729**	**0.7722**	**0.7562**	**2.83**
		F	T	0.7729	0.7511	0.7347	0.7339	0.7238	0.7151	6.67
		T	F	0.6882	0.7418	0.7586	0.7768	0.77	0.735	5.00
		T	T	0.6876	0.6477	0.6374	0.657	0.6538	0.6388	17.33
3	Decision Tree	**F**	**F**	**0.6706**	**0.7122**	**0.6691**	**0.7278**	**0.676**	**0.6647**	**14.33**
		F	T	0.6682	0.6206	0.5948	0.6016	0.6233	0.6189	23.50
		T	F	0.6545	0.6652	0.6728	0.6869	0.6896	0.6826	15.83
		T	T	0.655	0.6215	0.5963	0.623	0.6215	0.6027	23.83
4	Gradient Boosting	F	F	0.6565	0.7264	0.7061	0.7012	0.7093	0.6874	13.83
		F	T	0.6564	0.6027	0.6274	0.5833	0.6105	0.5547	25.00
		T	**F**	**0.682**	**0.7289**	**0.727**	**0.7252**	**0.7202**	**0.7037**	**11.00**
		T	T	0.6803	0.6612	0.6416	0.6736	0.6264	0.6429	17.33
5	KNN	F	F	0.5796	0.6262	0.6365	0.6384	0.6333	0.6084	22.67
		F	T	0.5796	0.571	0.5328	0.5279	0.5083	0.4882	30.50
		T	**F**	**0.6444**	**0.6414**	**0.6456**	**0.629**	**0.6336**	**0.5989**	**21.33**
		T	T	0.6432	0.6437	0.6404	0.6186	0.6011	0.6062	23.00
6	Logistic Regression	F	F	0.6101	0.6921	0.712	0.7224	0.7317	0.7003	13.67
		F	T	0.6101	0.5781	0.6003	0.6005	0.5821	0.5488	27.00
		T	**F**	**0.7412**	**0.7405**	**0.7329**	**0.7299**	**0.7362**	**0.7311**	**7.50**
		T	T	0.742	0.734	0.6971	0.7058	0.6813	0.6334	12.17
7	Random Forest	F	F	0.6726	0.7279	0.7464	0.7653	0.7735	0.7108	8.83
		F	T	0.6733	0.6098	0.6174	0.6191	0.6032	0.5802	23.50
		T	**F**	**0.6844**	**0.7417**	**0.7565**	**0.7698**	**0.7744**	**0.7334**	**6.17**
		T	T	0.6866	0.6472	0.6393	0.6582	0.6503	0.6415	17.50
8	SVM	F	F	0.5	0.5	0.5	0.548	0.54	0.5158	30.33
		F	T	0.5	0.5	0.5	0.5	0.5	0.5	31.33
		T	F	0.6939	0.6376	0.5878	0.5584	0.5349	0.5239	24.17
		T	**T**	**0.6928**	**0.6848**	**0.6118**	**0.5565**	**0.5409**	**0.5202**	**22.33**

- Tree-based methods prove to be the most successful, including Random Forest, Gradient Boosting, Decision Tree, and the two methods mentioned in point one.
- Both SMOTE and Undersampling do not yield substantial improvements in the methods based on trees. But they do work quite well for the methods non-based on trees such as KNN, SVM, and LR.
- The results from the two datasets exhibit considerable similarity, confirming a degree of consistency across the methods. The same applies to the different thresholds of removing cases in each experiment.

Fig. 6 presents insightful findings, illustrating the average ranking position per method and technique across various thresholds for combining the two datasets. Notably, Balanced Bagging consistently achieves an impressive average position of 1.52, often securing the top spot, while Balanced RF follows closely with 5.58, demonstrating robust performance even without SMOTE or Undersampling (US). It's noteworthy that Balanced Bagging outperforms Balanced RF in less than half the time, emphasizing its efficiency.

Both Random Forest and Gradient Boosting, employing multiple trees and determining outcomes through the average of these trees, exhibit comparable results. Interestingly, undersampling and SMOTE prove effective for KNN, SVM, and Logistic Regression, demonstrating their utility in enhancing performance for certain methods.

In short, for methods relying on multiple trees, the use of undersampling or oversampling techniques seems unnecessary, as they inherently address imbalances by averaging decisions across trees. Conversely, methods lacking this inherent capability require additional preprocessing steps.

Moving to Fig. 7, which presents the average position per method across all thresholds for both datasets, tree-based methods, particularly those designed for balancing (Balanced Bagging with an average position of 5.54 and Balanced RF with 8.19), exhibit superior performance compared to non-tree-based methods. Furthermore, it indicates that SMOTE enhances the ranking for non-tree-based algorithms while adversely affecting tree-based ones. Similarly, undersampling proves beneficial for non-tree-based algorithms but hinders performance for tree-based methods.

To summarize, the best strategy is to use Balanced Bagging and Balanced RF without SMOTE or Undersampling techniques. These tree-based methods consistently outperform their counterparts across various thresholds and datasets.

Compared to other data and algorithm-level approaches, this proposal is a hybrid sampling technique supporting the trend of creating hybrid methods. These methods make the performance of classifiers more efficient, even with few samples.

Due to privacy concerns and ethical considerations, the datasets used in this study are not publicly available. Thus, it is impossible to compare our approach's performance directly with other methods not tested on these specific datasets.

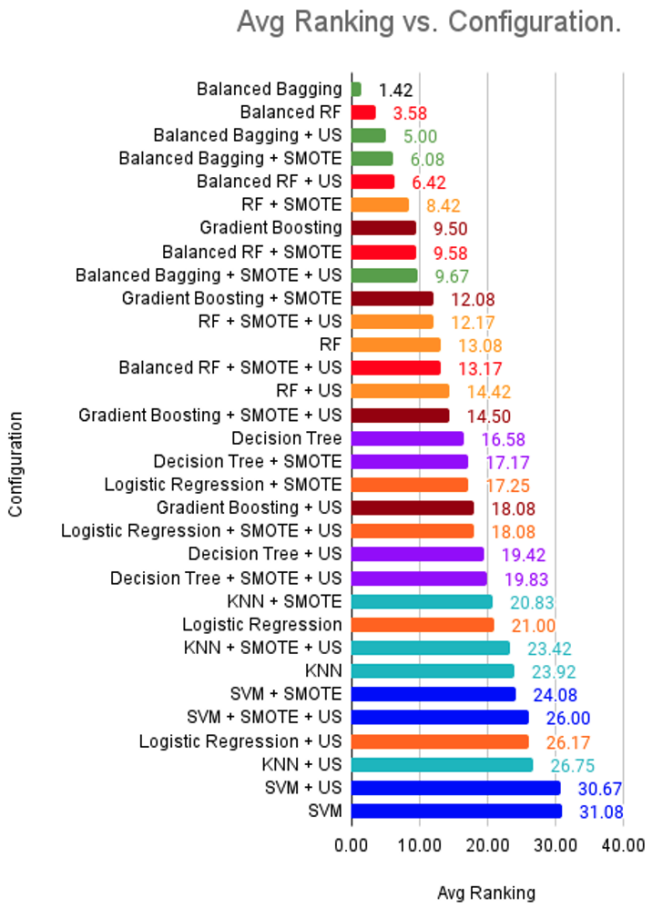

Fig. 6. Ranking positions for the optimal configuration of each supervised algorithm across 32 settings.

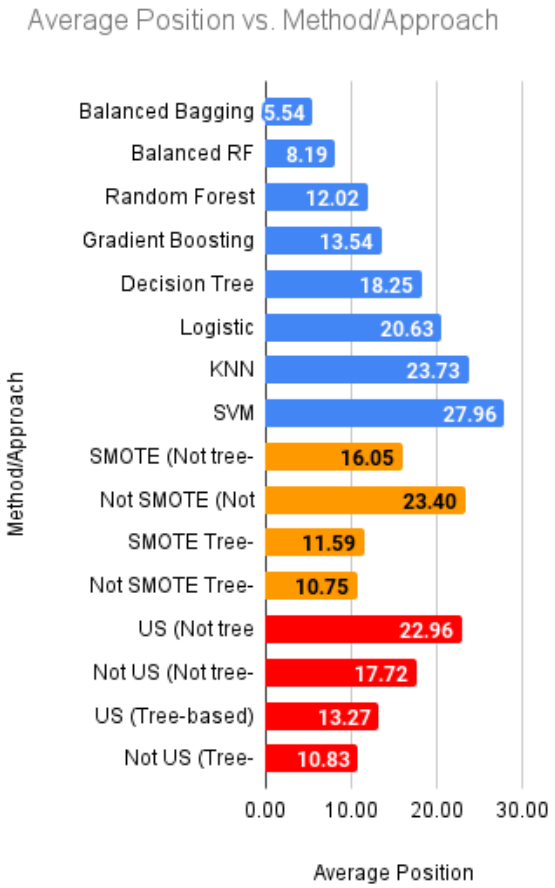

Fig. 7. Average rankings were averaged across different classification methods. They were also divided based on the use of SMOTE and undersampling, depending on whether they were tree-based.

However, we have chosen evaluation metrics carefully and compared our results with relevant benchmarks and state-of-the-art approaches on publicly available datasets.

VI. Conclusion

In this study, we aimed to enhance the accuracy and reliability of medical data classification models, ensuring both effectiveness and comprehensibility. By addressing class imbalance in medical datasets, we introduced a hybrid approach combining oversampling and undersampling techniques within ensemble classifiers. Our methodology sought to create balanced datasets and mitigate class imbalance effects, improving model performance and interpretability.

A. Conclusions

Our approach utilized random oversampling and SMOTE to generate balanced datasets, alongside transforming binary vectors to reduce bias in categorical variables. We also applied undersampling to the majority class to balance the number of instances between classes.

We assessed the performance of eight well-established ML classification methods using balanced accuracy, a metric that accounts for performance across both classes. Results highlighted that Balanced bagging and Balanced RF consistently outperformed others, demonstrating superior robustness in handling class imbalance and achieving high BA scores. Among tree-based methods, RF and GB exhibited the most success. Conversely, non-tree-based methods such as SVM, LR, and k-NN showed improved performance when combined with techniques like SMOTE and undersampling. The findings are further illustrated through figures showing average ranking positions, emphasizing the effectiveness of Balanced Bagging and the overall superiority of tree-based methods.

B. Future Work

Future research should focus on refining the hybrid sampling techniques to enhance adaptability across diverse medical datasets. Investigating the integration of emerging ML models and exploring their impact on class-imbalanced healthcare data could provide valuable insights. Extending the research to include more diverse medical conditions and datasets would contribute to a broader understanding of the proposed approach's generalizability. Addressing model interpretability and clinical relevance remains crucial. Further investigation into the computational efficiency of alternative sampling meth-

ods and their implementation in real-time healthcare applications is also recommended.

GLOSSARY

- **AMD**: Age-related macular Degeneration is a medical condition affecting the retina.
- **Balanced Accuracy**: This metric evaluates classification performance by averaging sensitivity and specificity to account for class imbalance.
- **Balanced Bagging**: Class imbalance is addressed by combining multiple models with resampling techniques using this algorithm.
- **Balanced Random Forest**: Effective handling of imbalanced datasets is achieved by adjusting the training of random forests with this algorithm.
- **Hybrid Sampling Technique**: Various sampling techniques are combined with this method to manage class imbalance in datasets.
- **Imbalanced Class Distributions**: A scenario where some classes are significantly underrepresented compared to others is described by this term.
- **Modified Dynamic Gini Index (MDGI)**: Assessing variable importance is done by measuring its impact on reducing impurity in predictive models with this metric.
- **Oversampling**: This technique involves adding additional samples to the minority class to balance class distributions.
- **Preeclampsia**: Characterized by high blood pressure and potential damage to organs, this pregnancy complication requires careful monitoring.
- **Recursive Feature Elimination (RFE)**: Improvement in model performance is achieved through iterative removal of the least important features in this feature selection process.
- **SMOTE (Synthetic Minority Over-sampling Technique)**: Generating synthetic examples for the minority class is done by creating new instances in feature space with this oversampling method.
- **Undersampling**: Removing samples from the majority class to balance the dataset is the focus of this technique.

REFERENCES

[1] S. Makridakis, "The forthcoming artificial intelligence (ai) revolution: Its impact on society and firms," *Futures*, vol. 90, pp. 46–60, 2017, doi 10.1016/j.futures.2017.03.006.

[2] V. Noorbakhsh-Sabet, N. Zand, Y. Zhang, and A. Abedi, "Artificial intelligence transforms the future of health care," *The American Journal of Medicine*, pp. 795–801, 2019, doi 10.1016/j.amjmed.2019.01.017.

[3] M. Khushi, K. Shaukat, T. M. Alam, I. A. Hameed, S. Uddin, S. Luo, X. Yang, and M. C. Reyes, "A comparative performance analysis of data resampling methods on imbalance medical data," *IEEE Access*, vol. 9, pp. 109 960–109 975, 2021, doi 10.1109/ACCESS.2021.3102399.

[4] Centro del Conocimiento Bioético, "Comisión nacional de bioética :: México," 2015.

[5] M. Bach, A. Werner, J. Żywiec, and W. Pluskiewicz, "The study of under- and over-sampling methods' utility in the analysis of highly imbalanced data on osteoporosis," *Medical Science Monitor*, pp. 174–190, 2017, doi 10.1016/j.ins.2016.09.038.

[6] K. T. M. Johnson, Justin M., "Survey on deep learning with class imbalance," *Journal of Big Data*, 2019, doi 10.1186/s40537-019-0192-5.

[7] M. P. Reddy, S. J. Fox, and Purohit, "Artificial intelligence-enabled healthcare delivery," *Journal of the Royal Society of Medicine*, vol. 112, no. 1, pp. 22–28, 2019, doi 10.1177/01410768188155.

[8] J. Xiao, C. Choi, and J. Sun, "Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review," *Journal of Medical Systems*, 2018, doi 10.1093/jamia/ocy068.

[9] A. Fernández, V. López, M. Galar, M. J. Del Jesus, and F. Herrera, "Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches," *Knowledge-Based Systems*, vol. 42, pp. 97–110, 2013, doi 10.1016/j.knosys.2013.01.018.

[10] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016, doi 10.1007/s13748-016-0094-0.

[11] W. L. Wong, X. Su, X. Li, C. M. G. Cheung, R. Klein, C.-Y. Cheng, and T. Y. Wong, "Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis," *The Lancet Global Health*, vol. 2, no. 2, pp. e106–e116, 2014.

[12] ——, "Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis," *The Lancet. Global health*, vol. 2, no. 2, pp. e106–16, February 2014, doi 10.1016/S2214-109X(13)70145-1.

[13] INEGI, "Banco de indicadores - ixtacamaxtitlán," 2016.

[14] T. A. Sivakumaran, R. P. Igo, J. M. Kidd, A. Itsara, L. J. Kopplin, W. Chen, S. A. Hagstrom, N. S. Peachey, P. J. Francis, M. L. Klein, E. Y. Chew, V. L. Ramprasad, W. T. Tay, P. Mitchell, M. Seielstad, D. E. Stambolian, A. O. Edwards, K. E. Lee, D. V. Leontiev, G. Jun, Y. Wang, L. Tian, F. Qiu, A. K. Henning, T. LaFramboise, P. Sen, M. Aarthi, R. George, R. Raman, M. K. Das, L. Vijaya, G. Kumaramanickavel, T. Y. Wong, A. Swaroop, G. R. Abecasis, R. Klein, B. E. K. Klein, D. A. Nickerson, E. E. Eichler, and S. K. Iyengar, "A 32 kb critical region excluding y402h in cfh mediates risk for age-related macular degeneration," *PLoS ONE*, vol. 6, no. 10, 2011, doi 10.1371/journal.pone.0209943.

[15] E. M. Stone, A. J. Aldave, A. V. Drack, M. W. MacCumber, V. C. Sheffield, E. Traboulsi, and R. G. Weleber, "Recommendations for genetic testing of inherited eye diseases: Report of the american academy of ophthalmology task force on genetic testing," *Ophthalmology*, vol. 119, no. 11, pp. 2408–2410, November 2012, doi 10.1016/j.ophtha.2012.05.047.

[16] L. Hindorff, J. MacArthur, J. HA, H. PN, K. AK, and M. TA, "Catalog of published genome-wide association studies - national human genome research institute (nhgri)," 2014.

[17] R. T. Yanagihara, C. S. Lee, D. S. W. Ting, and A. Y. Lee, "Methodological challenges of deep learning in optical coherence tomography for retinal diseases: a review," *Translational Vision Science & Technology*, vol. 9, no. 2, pp. 11–11, 2020, doi 10.1167/tvst.9.2.11.

[18] T. M. Alam, K. Shaukat, I. A. Hameed, W. A. Khan, M. U. Sarwar, F. Iqbal, and S. Luo, "A novel framework for prognostic factors identification of malignant mesothelioma through association rule mining," *Biomedical Signal Processing and Control*, vol. 68, p. 102726, 2021, doi 10.1016/j.bspc.2021.102726.

[19] P. Cacheiro Martínez, J. M. Ordovás, and D. Corella, "Métodos de selección de variables en estudios de asociación genética. aplicación a un estudio de genes candidatos en enfermedad de parkinson," Universidad de Santiago de Compostela, Coruña, España, Tech. Rep., 2011.

[20] R. Iniesta, E. Guinó, and V. Moreno, "Análisis estadístico de polimorfismos genéticos en estudios epidemiológicos," *Gaceta Sanitaria*, vol. 19, no. 4, pp. 333–341, 2005.

[21] M. Zhang and P. N. Baird, "A decade of age-related macular degeneration risk models: What have we learned from them and where are we going?" *Ophthalmic Genetics*, vol. 00, pp. 1–7, November 2016, doi 10.1080/13816810.2016.1227451.

[22] L. Sobrin and J. M. Seddon, "Nature and nurture-genes and environment-predict onset and progression of macular degeneration," *Progress in retinal and eye research*, vol. 40, pp. 1–15, 2014, doi 10.1016/j.preteyeres.2013.12.004.

[23] C. Castaneda, K. Nalley, C. Mannion, P. Bhattacharyya, P. Blake, A. Pecora, A. Goy, and K. S. Suh, "Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine," *Journal of Clinical Bioinformatics*, vol. 5, no. 1, p. 4, 2015, doi 10.1186/s13336-015-0019-3.

[24] A. Martínez-Velasco, L. Martínez-Villaseñor, A. C. Perez-Ortiz, J. C. Zenteno, A. B. Luna-Angulo, A. R. Villa-Romero, A. Rendon, F. J. Estrada, L. Martínez-Villasenor, L. Miralles-Pechuan, A. Rendon, and F. J. Estrada-Mena, "Cfh and htra1 genes associated with amd in

mexican population," *Investigative Ophthalmology & Visual Science*, vol. 58, no. 8, p. 2268, 2017, doi 10.13140/RG.2.2.10175.61609.

[25] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, and V. Robles, "Machine learning in bioinformatics," *Briefings in Bioinformatics*, vol. 7, no. 1, pp. 86–112, March 2006, doi 10.1016/B978-0-323-89775-4.00020-1.

[26] K. L. Spencer, L. M. Olson, N. Schnetz-Boutaud, P. Gallins, A. Agarwal, A. Iannaccone, S. B. Kritchevsky, M. Garcia, M. A. Nalls, A. B. Newman, W. K. Scott, M. A. Pericak-Vance, and J. L. Haines, "Using genetic variation and environmental risk factor data to identify individuals at high risk for age-related macular degeneration," *PLoS ONE*, vol. 6, no. 3, p. e17784, March 2011, doi 10.1371/journal.pone.0017784.

[27] R. Jiang, W. Tang, X. Wu, and W. Fu, "A random forest approach to the detection of epistatic interactions in case-control studies," *BMC Bioinformatics*, vol. 10, no. Suppl 1, p. S65, January 2009, doi 10.1186/1471-2105-10-S1-S65.

[28] B. Gold, J. C. J. E. Merriam, J. Zernant, L. S. Hancox, A. J. Taiber, K. Gehrs, K. Cramer, J. Neel, J. Bergeron, G. R. Barile, R. T. Smith, G. S. Hageman, M. Dean, R. Allikmets, S. Chang, L. A. Yannuzzi, I. Barbazetto, L. E. Lerner, S. Russell, J. Hoballah, J. Hageman, and H. Stockman, "Variation in factor b (bf) and complement component 2 (c2) genes is associated with age-related macular degeneration," *Nature Genetics*, vol. 38, no. 4, pp. 458–462, April 2006, doi 10.1038/ng1750.

[29] X. Chen, C.-T. Liu, M. Zhang, and H. Zhang, "A forest-based approach to identifying gene and gene-gene interactions," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 49, pp. 19 199–19 203, 2007, doi 10.1073/pnas.0709868104.

[30] A. Celebiler, H. Seker, B. YÜKSEL, A. Orun, S. Bilgili, and M. B. Karaca, "Discovery of the connection among age-related macular degeneration, mthfr c677t and pai 1 4g/5g gene polymorphisms, and body mass index by means of bayesian inference methods," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 21, no. 7, pp. 2062–2078, 2013, doi 10.3906/elk-1111-21.

[31] P. Fraccaro, M. Nicolo, M. Bonetto, M. Giacomini, P. Weller, C. E. Traverso, M. Prosperi, D. OSullivan, and D. OSullivan, "Combining macula clinical signs and patient characteristics for age-related macular degeneration diagnosis: a machine learning approach," *BMC Ophthalmology*, vol. 15, p. 10, January 2015, doi 10.1186/1471-2415-15-10.

[32] S. Krishnaiah, B. Surampudi, and J. Keeffe, "Modeling the risk of age-related macular degeneration and its predictive comparisons in a population in south india," *International Journal of Community Medicine and Public Health*, vol. 2, no. 2, p. 137, 2015, doi 10.5455/2394-6040.ijcmph20150514.

[33] K. E. Walker, *Evaluation of Children's Futures: Improving Health and Development Outcomes for Children in Trenton, New Jersey, 2001-2005*, 2008, inter-University Consortium for Political and Social Research.

[34] M. Sircar, R. Thadhani, and S. A. Karumanchi, "Pathogenesis of preeclampsia," *Current Opinion in Nephrology and Hypertension*, vol. 24, no. 2, pp. 131–138, 2015, doi 10.1097/MNH.0000000000000105.

[35] ACOG Committee on Practice Bulletins–Obstetrics, "ACOG practice bulletin. Diagnosis and management of preeclampsia and eclampsia. Number 33, January 2002," *Obstetrics and Gynecology*, vol. 99, no. 1, pp. 159–167, January 2002, doi 10.1016/s0029-7844(01)01747-1.

[36] A. R. Vest and L. S. Cho, "Hypertension in pregnancy," *Current atherosclerosis reports*, vol. 16, pp. 1–11, 2014, doi 10.1007/s11883-013-0395-8.

[37] M. A. Kohn, C. R. Carpenter, and T. B. Newman, "Understanding the direction of bias in studies of diagnostic test accuracy," *Academic Emergency Medicine*, vol. 20, no. 11, pp. 1194–1206, 2013, doi 10.1111/acem.12255.

[38] P. M. M. Bossuyt, "Clinical validity: Defining biomarker performance," *Scandinavian Journal of Clinical and Laboratory Investigation*, vol. 70, no. sup242, pp. 46–52, January 2010, doi 10.3109/00365513.2010.493383.

[39] L. C. Kenny, W. B. Dunn, D. I. Ellis, J. Myers, P. N. Baker, and D. B. Kell, "Novel biomarkers for pre-eclampsia detected using metabolomics and machine learning," *Metabolomics*, vol. 1, no. 3, pp. 227–234, 2005, doi 10.1007/s11306-005-0003-1.

[40] C. K. Neocleous, P. Anastasopoulos, K. H. Nikolaides, C. N. Schizas, and K. C. Neokleous, "Neural networks to estimate the risk for preeclampsia occurrence," in *Proceedings of the International Joint Conference on Neural Networks*, 2009, pp. 2221–2225, doi 10.1109/IJCNN.2009.5178820.

[41] M. Espinilla, J. Medina, A.-L. García-Fernández, S. Campaña, and J. Londoño, "Fuzzy intelligent system for patients with preeclampsia in wearable devices," *Mobile Information Systems*, pp. 1–10, October 2017, doi 10.1155/2017/7838464.

[42] M. Velikova, J. T. Van Scheltinga, P. J. Lucas, and M. Spaanderman, "Exploiting causal functional relationships in bayesian network modelling for personalised healthcare," *International Journal of Approximate Reasoning*, vol. 55, no. 1 PART 1, pp. 59–73, 2014, doi 10.1016/j.ijar.2013.03.016.

[43] E. Tejera, M. Jose Areias, A. Rodrigues, A. Rama, J. Manuel Nieto-Villar, and I. Rebelo, "Artificial neural network for normal, hypertensive, and preeclamptic pregnancy classification using maternal heart rate variability indexes," *Journal of Maternal-Fetal and Neonatal Medicine*, vol. 24, no. 9, pp. 1147–1151, 2011, doi 10.3109/14767058.2010.545916.

[44] P. M. Villa, P. Marttinen, J. Gillberg, A. I. Lokki, K. Majander, M. R. Ordén, P. Taipale, A. Pesonen, K. Räikkönen, E. Hämäläinen, E. Kajantie, and H. Laivuori, "Cluster analysis to estimate the risk of preeclampsia in the high-risk prediction and prevention of preeclampsia and intrauterine growth restriction (predo) study," *PLoS ONE*, vol. 12, no. 3, pp. 1–14, 2017, doi 10.1371/journal.pone.0174399.

[45] M. W. Moreira, J. J. Rodrigues, A. M. Oliveira, R. F. Ramos, and K. Saleem, "A preeclampsia diagnosis approach using bayesian networks," in *2016 IEEE International Conference on Communications (ICC)*, 2016, pp. 1–5.

[46] P. Fergus, C. C. Montanez, B. Abdulaimma, P. Lisboa, C. Chalmers, and B. Pineles, "Utilizing deep learning and genome-wide association studies for epistatic-driven preterm birth classification in african-american women," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 17, no. 2, pp. 668–678, 2018.

[47] B. Cox, P. Sharma, A. I. Evangelou, K. Whiteley, V. Ignatchenko, A. Ignatchenko, D. Baczyk, M. Czikk, J. Kingdom, J. Rossant, A. O. Gramolini, S. L. Adamson, and T. Kislinger, "Translational analysis of mouse and human placental protein and mrna reveals distinct molecular pathologies in human preeclampsia," *Molecular & Cellular Proteomics*, vol. 10, no. 12, p. M111.012526, 2011, doi 10.1074/mcp.M111.012526.

[48] R. Mehta, N. Bhatt, and A. Ganatra, "A survey on data mining technologies for decision support system of maternal care domain," *International Journal of Computer Applications*, vol. 138, no. 10, pp. 975–8887, 2016, doi 10.5120/ijca2016908965.

[49] G. Kovács, "Smote-variants: A python implementation of 85 minority oversampling techniques," *Neurocomputing*, vol. 366, pp. 352–354, November 2019, doi 10.1016/j.neucom.2019.06.100.

[50] J. Luengo, A. Fernández, S. García, and F. Herrera, "Addressing data complexity for imbalanced data sets: analysis of smote-based oversampling and evolutionary undersampling," *Soft Computing*, vol. 15, no. 10, pp. 1909–1936, 2011, doi 10.1007/s00500-010-0625-8.

[51] G. E. A. P. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, p. 20, June 2004, doi 10.1145/1007730.1007735.

[52] B. Zhu, Z. Gao, J. Zhao, and S. K. vanden Broucke, "Iric: An r library for binary imbalanced classification," *SoftwareX*, vol. 10, p. 100341, 2019.

[53] K.-J. Wang, A. M. Adrian, K.-H. Chen, and K.-M. Wang, "A hybrid classifier combining borderline-smote with airs algorithm for estimating brain metastasis from lung cancer: A case study in taiwan," *Computer methods and programs in biomedicine*, vol. 119, no. 2, pp. 63–76, 2015, doi 10.1016/j.cmpb.2015.03.003.

[54] X. Y. Liu and Z. H. Zhou, "The influence of class imbalance on cost-sensitive learning: An empirical study," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2006, pp. 970–974, doi 10.1109/ICDM.2006.158.

[55] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 8, pp. 3573–3587, August 2018.

[56] G. M. Weiss and F. Provost, "Learning when training data are costly: The effect of class distribution on tree induction," *Journal of Artificial Intelligence Research*, vol. 19, pp. 315–354, 2007, doi 10.1613/jair.1199.

[57] K. McCarthy, B. Zabar, and G. Weiss, "Does cost-sensitive learning beat sampling for classifying rare classes?" in *Proceedings of the 1st International Workshop on Utility-Based Data Mining, UBDM '05*, 2005, pp. 69–77, doi 10.1145/1089827.1089836.

[58] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002, doi 10.3233/IDA-2002-6504.

[59] M. D. Alanis Tamez, "Prediagnóstico de enfermedades crónicas mediante algoritmos de cómputo inteligente," Ph.D. dissertation, CIC, IPN, 2018, doi 10.13053/cys-24-3-3492.

[60] B. U. A. M. Moreno, "Sistema de clasificación paralelo basado en un ensamble de tipo mezcla de expertos," Ph.D. dissertation, Universidad Autónoma Metropolitana, 2017.

[61] Y. Xia, K. Chen, and Y. Yang, "Multi-label classification with weighted classifier selection and stacked ensemble," *Information Sciences*, 2020, doi 10.1016/j.ins.2020.06.017.

[62] J. Diez-Pastor, J. J. Rodríguez, C. García-Osorio, and L. Kuncheva, "Diversity techniques improve the performance of the best imbalance learning ensembles," *Information Sciences*, vol. 325, pp. 98–117, 2015, doi 10.1016/j.ins.2015.07.025.

[63] L. J. Mena, E. E. Orozco, V. G. Felix, R. Ostos, J. Melgarejo, and G. E. Maestre, "Machine learning approach to extract diagnostic and prognostic thresholds: Application in prognosis of cardiovascular mortality," *Computational and Mathematical Methods in Medicine*, vol. 2012, 2012, doi 10.1155/2012/750151.

[64] J. Li, Y. Liu, and Q. Li, "Intelligent fault diagnosis of rolling bearings under imbalanced data conditions using attention-based deep learning method," *Measurement*, vol. 189, p. 110500, 2022, doi 10.1088/1742-6596/2369/1/012001.

[65] L. I. Santos, M. O. Camargos, M. F. S. V. D'Angelo, J. B. Mendes, E. E. C. de Medeiros, A. L. S. Guimarães, and R. M. Palhares, "Decision tree and artificial immune systems for stroke prediction in imbalanced data," *Expert Systems with Applications*, vol. 191, p. 116221, 2022.

[66] X. Peng, X. Jin, S. Duan, and C. Sankavaram, "Active learning-assisted semi-supervised learning for fault detection and diagnostics with imbalanced dataset," *IISE Transactions*, vol. 55, no. 7, pp. 672–686, 2023, doi 10.1080/24725854.2022.2074579.

[67] H. Ullah, M. B. B. Heyat, F. Akhtar, A. Y. Muaad, C. C. Ukwuoma, M. Bilal, M. H. Miraz, M. A. S. Bhuiyan, K. Wu, R. Damaševičius *et al.*, "An automatic premature ventricular contraction recognition system based on imbalanced dataset and pre-trained residual network using transfer learning on ecg signal," *Diagnostics*, vol. 13, no. 1, p. 87, 2023, doi 10.3390/diagnostics13010087.

[68] M. Mazur-Milecka, N. Kowalczyk, K. Jaguszewska, D. Zamkowska, D. Wójcik, K. Preis, H. Skov, S. Wagner, P. Sandager, M. Sobotka *et al.*, "Preeclampsia risk prediction using machine learning methods trained on synthetic data," in *Polish Conference on Biocybernetics and Biomedical Engineering*. Springer, 2023, pp. 267–281, doi 10.1007/978-3-031-38430-1-21.

[69] V. P. Kovacheva, B. W. Eberhard, R. Y. Cohen, M. Maher, R. Saxena, and K. J. Gray, "Prediction of preeclampsia from clinical and genetic risk factors in early and late pregnancy using machine learning and polygenic risk scores," *MedRxiv*, pp. 2023–02, 2023, doi 10.1161/HYPERTENSIONAHA.123.21053.

[70] A. R. Chłopowiec, K. Karanowski, T. Skrzypczak, M. Grzesiuk, A. B. Chłopowiec, and M. Tabakov, "Counteracting data bias and class imbalance—towards a useful and reliable retinal disease recognition system," *Diagnostics*, vol. 13, no. 11, p. 1904, 2023, doi 10.3390/diagnostics13111904.

[71] Y. Xie, Q. Wan, H. Xie, Y. Xu, T. Wang, S. Wang, and B. Lei, "Fundus image-label pairs synthesis and retinopathy screening via gans with class-imbalanced semi-supervised learning," *IEEE Transactions on Medical Imaging*, 2023, doi.

[72] Y. A. Veturi, W. Woof, T. Lazebnik, I. Moghul, P. Woodward-Court, S. K. Wagner, T. A. C. de Guimarães, M. D. Varela, B. Liefers, P. J. Patel *et al.*, "Syntheye: Investigating the impact of synthetic data on artificial intelligence-assisted gene diagnosis of inherited retinal disease," *Ophthalmology Science*, vol. 3, no. 2, p. 100258, 2023, doi.

[73] K.-J. Wang, A. M. Adrian, K.-H. Chen, and K.-M. Wang, "A hybrid classifier combining borderline-smote with airs algorithm for estimating brain metastasis from lung cancer: a case study in taiwan," *Computer Methods and Programs in Biomedicine*, vol. 119, no. 2, p. 63—76, April 2015, doi 10.1016/j.cmpb.2015.03.003.

[74] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, January 2002, doi 10.5555/1622407.1622416.

[75] V. Palodeto, H. Terenzi, and J. L. B. Marques, "Training neural networks for protein secondary structure prediction: the effects of imbalanced data set," in *International Conference on Intelligent Computing*, 2009, pp. 258–265, doi 10.1007/978-3-642-04020-7-28.

[76] R. Blagus and L. Lusa, "Smote for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 14, 2013, doi 10.1186/1471-2105-14-106.

[77] F. Rodríguez Torres, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "Smote-d a deterministic version of smote," in *Mexican Conference on Pattern Recognition*, 2016, pp. 177–188, doi 10.1007/978-3-319-39393-3-18.

[78] S. Vluymans, "Learning from imbalanced data," *Studies in Computational Intelligence*, vol. 807, pp. 81–110, 2019, doi 10.1007/978-3-030-04663-7-4.

[79] A. J. Mohammed, M. M. Hassan, and D. H. Kadir, "Improving classification performance for a novel imbalanced medical dataset using smote method," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 3, pp. 3161–3172, May 2020, doi 10.30534/ijatcse/2020/104932020.

[80] Z. Yao, Y. Yuan, Z. Shi, W. Mao, G. Zhu, G. Zhang, and Z. Wang, "Funswin: A deep learning method to analysis diabetic retinopathy grade and macular edema risk based on fundus images," *Frontiers in Physiology*, vol. 13, p. 961386, 2022.

[81] M. S. Khan, N. Tafshir, K. N. Alam, A. R. Dhruba, M. M. Khan, A. A. Albraikan, F. A. Almalki *et al.*, "Deep learning for ocular disease recognition: an inner-class balance," *Computational Intelligence and Neuroscience*, vol. 2022, 2022, doi 10.1155/2022/5007111.

[82] T. M. Alam, K. Shaukat, W. A. Khan, I. A. Hameed, L. A. Almuqren, M. A. Raza, M. Aslam, and S. Luo, "An efficient deep learning-based skin cancer classifier for an imbalanced dataset," *Diagnostics*, vol. 12, no. 9, p. 2115, 2022, doi 10.3390/diagnostics12092115.

[83] A. M. Sowjanya and O. Mrudula, "Effective treatment of imbalanced datasets in health care using modified smote coupled with stacked deep learning algorithms," *Applied Nanoscience*, vol. 13, no. 3, pp. 1829–1840, 2023, doi 10.1007/s13204-021-02063-4.

[84] K. Koc, Ö. Ekmekcioğlu, and A. P. Gurgun, "Prediction of construction accident outcomes based on an imbalanced dataset through integrated resampling techniques and machine learning methods," *Engineering, Construction and Architectural Management*, 2022, doi 10.1108/ECAM-04-2022-0305.

[85] X.-w. Chen and M. Wasikowski, "Fast: a roc-based feature selection metric for small samples and imbalanced data classification problems," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 124–132, doi 10.1145/1401890.1401910.

[86] Z. P. Agusta *et al.*, "Modified balanced random forest for improving imbalanced data prediction," *International Journal of Advances in Intelligent Informatics*, vol. 5, no. 1, pp. 58–65, 2019, doi 10.26555/ijain.v5il.255.

[87] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001, doi 10.1007/978-1-4419-9326-7-5.

[88] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001, doi 10.1214/aos/1013203451.

[89] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, pp. 81–106, 1986, doi 10.1007/BF00116251.

[90] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 20, no. 2, pp. 215–232, 1958, doi 10.1111/j.2517-6161.1959.tb00334.x.

[91] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967, doi 10.1109/TIT.1967.1053964.

[92] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995, doi 10.1007/BF00994018.

[93] L. I. Brigato, Lorenzo, "A close look at deep learning with small data," *IEEE, 25th international conference on pattern recognition*, 2021, 10.1109/ICPR48806.2021.9412740.

[94] M. L. Calle and V. Urrea, "Stability of random forest importance measures," *Briefings in bioinformatics*, vol. 12, no. 1, pp. 86–89, 2011, doi 10.1093/bib/bbq011.

[95] M. S. Shelke, P. R. Deshmukh, and V. K. Shandilya, "A review on imbalanced data handling using undersampling and oversampling technique," *Int. J. Recent Trends Eng. Res*, vol. 3, no. 4, pp. 444–449, 2017.

Antonieta Martínez-Velasco is a Research Professor in the School of Engineering at Universidad Panamericana, Mexico. She is a member of the Mexican National System of Researchers (CONACYT). She has a Ph.D. in Engineering from the Universidad Panamericana. Her specialization is artificial intelligence, applied to engineering, health sciences, and social sciences.

Lourdes Martínez-Villaseñor is a Full-time Professor in the School of Engineering at the Universidad Panamericana, Mexico, and head of the postgraduate academic area. She is a Computer Systems Engineer and a Doctor in Computational Sciences from Tecnológico de Monterrey, Mexico. She has the distinction of level 1 of the National System of Researchers of CONACYT. Her main research interests are artificial intelligence applied to healthcare systems and ethics for artificial intelligence.

Luis Miralles-Pechuán is a Lecturer at Technological University Dublin. He obtained his PhD and Bachelor in Computer Science at the University of Murcia (Spain). He worked as a full-time researcher/lecturer at the University Panamericana in Mexico for three years. He started a PhD in 2012 on creating new approaches within the Online Advertising world. During his PhD, he got familiar with ML and many papers on how to apply ML to online advertising. After finishing his PhD, he worked in postdoc levels I and II in CeADAR, University College Dublin, and there, he won the prize for supervising the best student paper at the Digital Forensic conference. His topic is applying Reinforcement Learning to fight the COVID-19 pandemic and plan the containing levels, considering public health and the economy. Lastly, he has expertise in human activity recognition and generalized zero-shot learning (GZSL) and applying machine learning to improve the accessibility of websites.