IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 March 2024, accepted 25 March 2024, date of publication 1 April 2024, date of current version 10 April 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3383837*

RESEARCH ARTICLE

# Evaluation of Machine Learning Techniques for Classifying and Balancing Data on an Unbalanced Mini-Mental State Examination Test Data Collection Applied in Chile

**PABLO ORMEÑO[1], GASTÓN MÁRQUEZ[2,4], AND CARLA TARAMASCO[3,4]**

[1]Escuela de Ingeniería y Negocios, Universidad de Viña del Mar, Viña del Mar 2580000, Chile
[2]Departamento de Ciencias de la Computación y Tecnologías de la Información, Universidad del Bío-Bío, Chillán 3780000, Chile
[3]Instituto de Tecnología para la Innovación en Salud y Bienestar, Facultad de Ingeniería, Universidad Andrés Bello, Viña del Mar, Valparaíso 2520000, Chile
[4]Millennium Nucleus on Sociomedicine, Santiago 8320000, Chile

Corresponding author: Pablo Ormeño (pablo.ormeno@uvm.cl)

This work was supported in part by FONDECYT Regular (Multimodal Machine Learning Approach for Detecting Pathological Activity Patterns in Elderlies) under Grant 1201787, and in part by the Project ANID Millennium Science Initiative Program Millennium Nucleus on Sociomedicine under Grant NCS2021_013.

**ABSTRACT** The Mini-Mental State Examination (MMSE) is the most widely used cognitive test for assessing whether suspected symptoms align with cognitive impairment or dementia. The results of this test are meaningful for clinicians but exhibit highly unbalanced distributions in studies and analyses regarding the classification of patients with cognitive impairment. This is a complex problem when a large number of MMSE tests are analysed. Therefore, data balancing and classification techniques are crucial to support decision-making in distinguishing patients with cognitive impairment in an effective and efficient manner. This study explores machine learning techniques for data balancing and classification using a real unbalanced dataset consisting of MMSE test responses collected from 103 elderly patients participating in a Chilean patient monitoring project. We used 8 data classification techniques and five data balancing techniques. We evaluated the performance of the techniques using the following metrics: sensitivity, specificity, F1-score, likelihood ratio (LR+ and LR-), diagnostic odds ratio (DOR), and the area under the ROC curve (AUC). From the set of data balancing and classification techniques used in this study, the results indicate that synthetic minority oversampling and random forest balancing techniques improve the accuracy of cognitive impairment diagnosis. The results obtained in this study support clinical decision-making regarding early classification or exclusion of older adult patients with suspected cognitive impairment.

**INDEX TERMS** Mini-mental state exam, machine learning, imbalanced data.

## I. INTRODUCTION

Decreased cognitive ability is one of the most important symptoms of Alzheimer's disease (AD). This decrease in cognitive ability may span several years, sometimes decades, starting from normal cognition (NC) and progressing to mild cognitive impairment (MCI). This transition causes a change in the condition of a patient from suspected AD to confirmed AD [1], [2]. In this scenario, early detection and precise diagnosis are important; nevertheless, detecting AD requires an extensive medical assessment, including patient history and physical and neurological examinations [3].

The diagnosis of dementia requires a cognitive assessment of brain functions such as attention, memory,

The associate editor coordinating the review of this manuscript and approving it for publication was Felix Albu.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

problem-solving, thinking, and mental abilities [4], [5]. While some of these assessments are invasive, costly, or stressful to the patient, an early and accurate diagnosis is important for controlling disease progression. In this regard, the Mini-Mental State Examination (MMSE) is a test in which a patient can score a maximum of 30 points [6], and it is frequently used to screen for dementia. The MMSE is widely used in clinical settings, particularly by health care professionals specializing in neurology, psychiatry, geriatrics, and primary care. This examination helps in identifying cognitive impairment, tracking changes in cognitive function over time, and determining the severity of cognitive decline.

Machine learning is a subset of artificial intelligence, computer science, and statistics where computational tasks are performed with algorithms that learn patterns from data to automatically make inferences. One of the main problems faced in machine learning is dataset class imbalance, which affects the quality and reliability of the solutions [7], [8], [9]. This is a very frequent problem because imbalanced data are a reality in almost all biological datasets [10]. The prediction of a rare condition is important, especially in the context of medical diagnosis, where promptly identifying a specific disease is critical, and the majority of patients are healthy [11]. This issue is very common in the field of diagnosis because of the numerous limitations in generating, managing, and acquiring new samples, especially clinical data, which heavily depend on a patient's willingness to release their data or participate in trials. Consequently, learning tasks can be difficult, and nonstandard machine learning methods are needed to achieve desirable results, especially in the presence of low-prevalence diseases or clinical conditions.

In this study, we explore the problem of imbalanced data in an MMSE dataset obtained as part of the EHomeSenior project [12]. We analyzed the utility of machine learning techniques in the context of an imbalanced MMSE dataset and evaluated the efficacy of sampling methodologies in improving the performance. The primary contribution of this research lies in its provision of a methodological analysis of machine learning techniques to address the challenge of class imbalance in diagnostic data pertaining to cognitive impairment.

The rest of this paper is structured as follows: Section II describes the methodology used in this study; Section III details the results; Section IV describes the discussion and key findings; Section V discuss related work; and Section VI concludes the research.

## II. METHODS

The methodology used in this study is illustrated in Figure 1.

The aim of our study is to explore the performance of machine learning techniques in terms of classifying previously balanced data in a real dataset consisting of results from an MMSE test administrated to older adults. The research question of our study is as follows: *Which data balancing and classification technique achieves better performance in diagnosing cognitive impairment on an unbalanced dataset?*

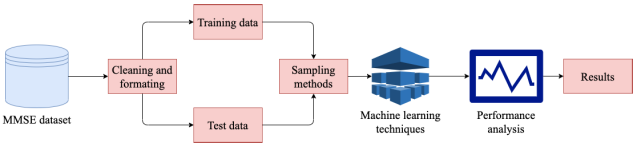


**FIGURE 1.** Proposed methodology.

### A. MMSE DATASET

The data used in this study are from the EHomeSenior project [12]. The EHomeSenior project employs a non-intrusive monitoring system that adapts to the daily routine of elderly individuals. This system continuously (24 h a day) monitors their daily activities to detect risk events (e.g., falls) and evaluate nocturia and actimetry, alerting family members about fall risks, carbon monoxide inhalation, early symptoms of degenerative diseases, and general safety hazards. This system uses the MMSE to measure the cognitive status of older adults. The process of administering the MMSE begins when sensors are installed in an older adult's home. A clinician then asks initial demographic questions to determine their habits, whether they use any assistive devices (cane, glasses, removable dental prosthetic), or if they have had surgery. Once the interview is completed, the MMSE is started, and the answers are sent to a server. In our study, the results form 103 participants were added to the dataset based on inclusion and exclusion criteria (see Table 1).

**TABLE 1.** Inclusion/exclusion criteria for MMSE patients.

| | |
|---|---|
| Inclusion Criteria | > 65 years old with perceived socioeconomic risk |
| | Resident of Valparaiso or San Antonio, Chile |
| | Lives alone |
| | Receives a housing subsidy from Housing and Urbanism Services |
| Exclusion Criteria | Suffers from dementia |
| | Suffers from substance or alcohol abuse |
| | Unable to answer quality of life questionnaires due to medical or psychiatric morbidity or terminal illness (life expectancy of less than 6 months) |
| | Pets within the residence (i.e., dogs, cats) |
| | Refusal to participate |

We defined these inclusion and exclusion criteria based on our previous experience [13], [14] performing an EQ5D [15] survey in research projects with older adults. As we have gained experience dealing with older adults, the inclusion and exclusion criteria defined in Table 1 have helped us obtain a meaningful set of older adults to conduct our study.

Concerning Table 1, patients with MMSE scores outside the range of 0 to 30 are eliminated.

### B. DATA CLEANING AND FORMATTING

The MMSE consists of 30 questions, some of which are simple and some of which are complex. This examination also includes tasks in certain areas, such as time and place orientation, word repetition, attention or calculation, language use, and motor skills. Table 2 summarizes the questions that were applied.

**TABLE 2. MMSE topics and quantity of questions.**

| Name | Number of questions |
|---|---|
| Temporal Orientation | 5 |
| Spatial Orientation | 5 |
| Registration | 3 |
| Attention and Calculation | 5 |
| Memory | 3 |
| Language | 2 |
| Repetition | 1 |
| Complex Commands | 6 |

Concerning the questions described in Table 2, Folstein et al. [16] explained the survey in detail, and Mitchell [17] discussed updates on diagnostic accuracy and its evolution. It is important to note that several resources on the Internet display MMSE surveys for the community. The original data distribution, consisting of 90% of NC and 10% of MCI, with a total of 103 older adults who participated in the eHomeSenior project, agreed to take the test. This distribution represents the sample of all older adults in our study. Given that the results obtained from the 103 participants are pioneering in Chile, we elected to analyse them as they were obtained in order to uncover more significant results in a clinical context. Therefore, we treated the problem as an imbalanced problem for binary classification. Consequently, to preprocess the answers, we used a binary vector that identifies every answer to every question. We added a 1 for every true answer and a 0 for every false answer.

### C. TRAINING AND TEST DATA

We trained the models using the NC and MCI classes to make predictions from MMSE data collection according to the features. To optimize the model testing process, we split the data into five folds. In each iteration, four different folds were used for training, and the remaining fold was used for testing. We used 5-fold cross validation to fit the models. Once this stage was completed, a dataset that the models had not seen was tested (this dataset is the same for all models). This dataset was used to report the model performance. On the other hand, we trained each classification model to learn the approximate function to classify the patients' MMSE values, thus minimizing the classification error. Each classification model was trained using the same training set (80% of the samples) and tested on the same test set (20% of the samples). The features were normalized using z scores. The evaluation metrics were generated over 20 runs, considering the random data distribution in each partition. The proposed approach was implemented in Python 3 using scikit-learn as the backend.

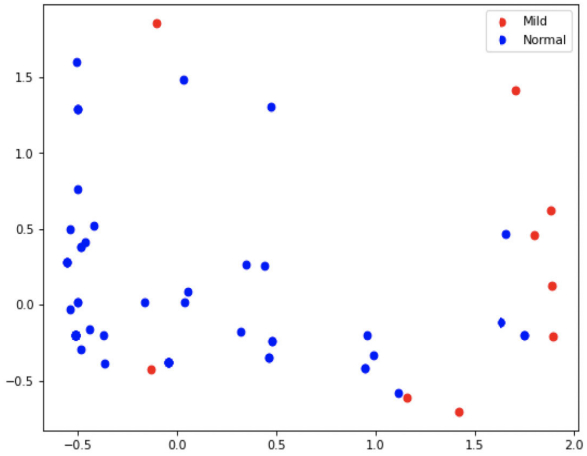


**FIGURE 2. Effect of the sampling methods on the original MMSE dataset using normal data distribution.**

### D. MACHINE LEARNING APPROACHES

We used machine learning techniques that have been used in different clinical studies analysing datasets, such as [10], [11], [18], [19], [20], and [21]. Table 3 summarizes the techniques used in this study.

Because the MMSE data are highly imbalanced, in this study, the amount of data available for MCI patients is small compared with that of the other class. In this regard, most classifiers are biased towards the class with more data (cognitively normal patients) and have poor classification rates for the class with less data. In other cases, the classifier may consider everything within a larger class and ignore smaller ones; this is a problem with multiclass data. Therefore, there are many techniques for handling imbalanced data problems. Often, these techniques classify the approaches as sampling methods (preprocessing) and cost-sensitive methods. Some sampling techniques are more accessible and do not require any specific information about the classification problem. In these cases, a new dataset is created to balance the classes, giving the classifiers a better opportunity to distinguish the decision boundary between them. Table 4 summarizes the methods used in this study.

Figure 2 represents the original balance of the data, where 9% of the data corresponds to normal cognition patients (normal) and 91% corresponds to mild cognitive impairment patients (mild). Additionally, Figures 3, 4, 5, 6 and 7 show the application of the techniques described in Table 4. Figure 2 to Figure 7 depict an $X$-$Y$ axis representation of the dataset, which is inherently multidimensional in nature. These graphics display a visualization of the data subsequent to the application of Principal Component Analysis (PCA) [33] and the reduction of all features to two dimensions. In some instances, the scale changes are a result of resampling

**TABLE 3.** Machine learning methods.

| Technique | Description |
|---|---|
| Naive Bayes (NB) [22] | NB counts the frequency of the values in a dataset and calculates the probabilities of each class. NB assumes that all attributes are conditionally independent, given the target value. |
| Support vector machine (SVM) [23] | The SVM is a supervised learning method for classification that works by finding a hyperplane (line in 2-dimension or plane in 3-dimension) capable of splitting the data into different classes. |
| K-nearest neighbour (KNN) [24] | This technique is based on the principle that instances from a dataset are similar when they have similar properties. In this case, when unclassified data appears, it is labelled according to its nearest neighbours. |
| Decision tree (DT) [25] | This technique classifies data according to their features, where each node represents a feature and each branch represents the value that a node can have. To achieve this objective, a binary tree is built based on a feature that better divides the data as a root node to classify data. |
| Random forest (RF) [26] | The RF is an ensemble of learning approaches that uses multiple nonpruned decision trees for classification. To generate the RF classifier, each decision tree is created from a subset of data features. |
| Logistic regression (LR) [27] | LR is a supervised classification technique that builds a regression model to predict the classes of a dataset by using the sigmoid function. As in linear models, in logistic regression, learning models compute a ponderated sum of the input features with bias. |
| XGBoost [28] | XGBoost is used on decision tree ensemble methods and learns from previous errors. Specifically, XGBoost uses the gradient of the loss function in the model for pseudoresidual calculations between the predicted and real labels. |
| Multilayer perceptron (MLP) [29] | An MLP is a connected neural network with a minimum of three layers of neurons (input, hidden, and output). The basic principle consists of a neuron representing a node and has an activation function, for example, the sigmoid function, which is activated according to the sum of the arriving weighted signals from previous layers. |

**TABLE 4.** Metrics used to compare algorithms.

| Name | Description |
|---|---|
| Random sampling [29] | There are two options: random undersampling (RUS) and random oversampling (ROS). In both the cases, the dataset is adjusted to create a new dataset with a more similar class distribution. In the case of undersampling, the majority of class instances are removed to balance the dataset. In contrast, in the ROS technique, the minority class is duplicated to achieve a better data distribution. |
| Synthetic minority oversampling technique (SMOTE) [30] | This method generates synthetic data in a dataset. To achieve balance in the minority dataset, SMOTE first randomly selects a minority class instance $M_a$. Then, the k-nearest neighbours of $M_a$ related to the minority class are identified. A second instance $M_b$ is selected from the set. In this way, $M_a$ and $M_b$ are connected, forming a line segment in the feature space. |
| Synthetic minority oversampling technique with Tomek link (SMOTETomek) [31] | SMOTE is improved by considering Tomek links, balancing the data, and creating a well-separated class instance. In this solution, every data instance forming a Tomek link is discarded, from minority and majority classes. |
| Adaptive synthetic sampling algorithm (ADASYN) [32] | This technique uses a density estimation metric as a criterion to create several synthetic samples of the minor class. Therefore, it is possible to balance minority and majority classes and create synthetic data. |

methods that generate new data samples that are proportionate to the existing data. When analysing Figures 2, 3, 4, 5, 6 and 7, it can be observed that the SMOTE, SMOTETomek, ADASYN, ROS and RUS techniques balance the distribution of the data.

### *E. EVALUATION METRICS*

When a classification task is performed, four possible outputs can be obtained. The classifier may correctly assign a sample as positive (target condition) or negative (without target condition), i.e., true positives (TP) and true negatives (TN), respectively. Alternatively, the classifier may make wrong predictions, where the true labels obtained from the gold standard contrast with the predicted labels, i.e., false-positives (FP) or false-negatives (FN), respectively. These results are summarized in the confusion matrix, as shown in Table 5.

**TABLE 5.** Confusion matrix of binary classification.

| Classifier | Mild Cognitive Subjects | Normal Subjects |
|---|---|---|
| Positive | TP | FP |
| Negative | FN | TN |

In the presence of imbalanced data, not all data can be used for diagnostic tasks [34]. The most well-known example is accuracy, which is widely used in classification. However, this metric does not necessarily reflect the biological significance of the results. For this reason, machine learning approaches should always be accompanied by expert decisions regarding the final result. In this paper, we focused

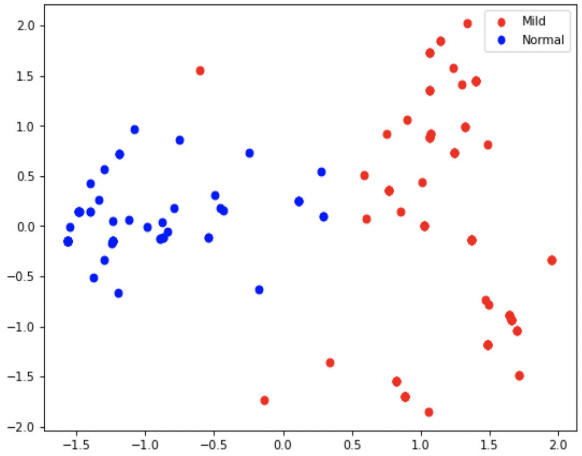


**FIGURE 3.** Effect of the sampling methods on the original MMSE dataset using SMOTE data distribution.

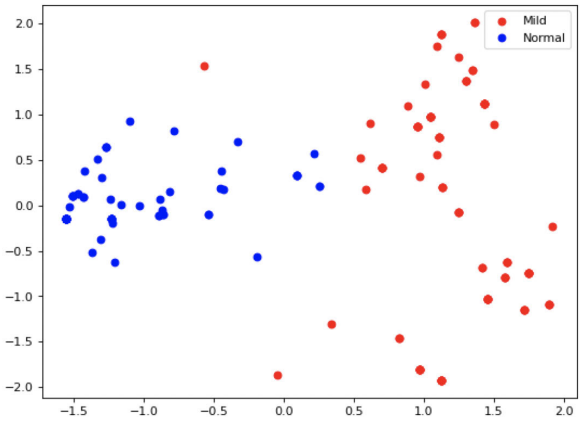


**FIGURE 4.** Effect of the sampling methods on the original MMSE dataset using SMOTETomek data distribution.

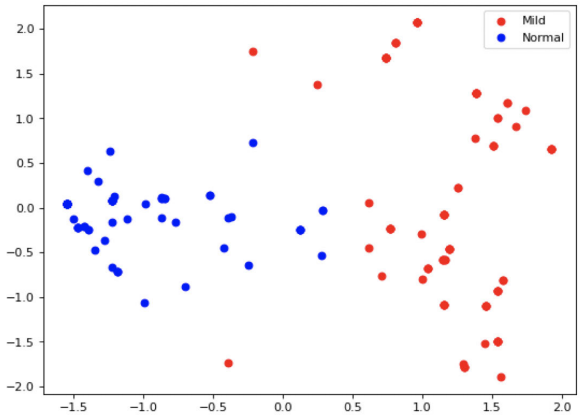


**FIGURE 5.** Effect of the sampling methods on the original MMSE dataset using ADASYN data distribution.

on seven distinct metrics commonly used in classification and diagnostic tasks, which can be used for imbalanced data [34], [35], [36] (see Table 6).

## III. RESULTS

In this section, we present the results of the experiments conducted to demonstrate the effects of resampling on

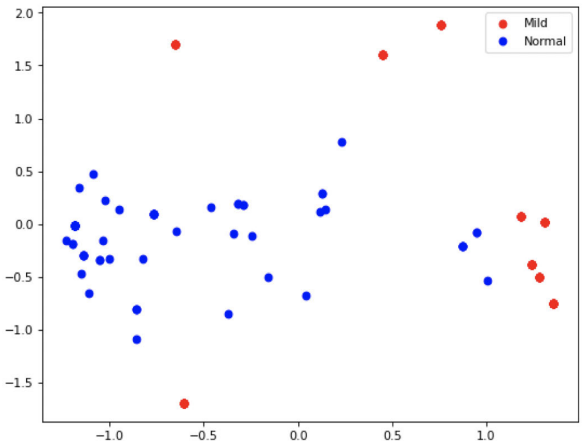


**FIGURE 6.** Effect of the sampling methods on the original MMSE dataset using ROS data distribution.

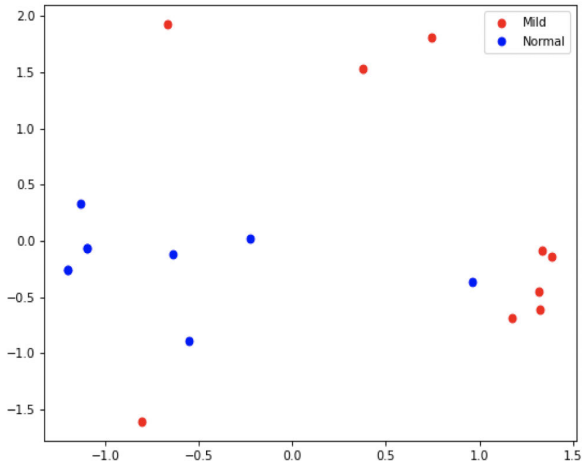


**FIGURE 7.** Effect of the sampling methods on the original MMSE dataset using RUS data distribution.

the original dataset. Table 7 shows the application of the techniques to the original unbalanced dataset, and Tables 8, 9, 10, 11, and 12 show the results of the application of the resampling techniques. The columns in each table show the performance of the classifiers (accuracy, F1, sensitivity, and specificity) along with the diagnostic metrics (LR+, LR−, AUC, and DOR). When evaluating the machine learning techniques, good performance was achieved in terms of accuracy, particularly for the RF, with the original dataset, as presented in Table 7.

However, as expected, the models exhibited poor performance in terms of F1-score and sensitivity. Notably, the RF also exhibited good discriminative power and diagnostic performance, as indicated by its high AUC and DOR. By applying ROS, as shown in Table 8, the class imbalance issue was effectively addressed, leading to improved model performance compared to that with the original dataset. The models achieved higher accuracies, F1-scores, and sensitivity values. This increase in accuracy indicates an overall improvement in the classification performance, whereas the improved F1-scores demonstrate a better balance between

**TABLE 6. Metrics used to compare algorithms.**

| Description | Formula | Metric |
|---|---|---|
| The proportion of correctly positive samples classified among all positive samples. | $\frac{TP}{TP+FN}$ | Sensitivity |
| The proportion of correctly classified negative samples among all negative samples (healthy control samples). | $\frac{TN}{FP+TN}$ | Specificity |
| The harmonic mean of the precision and recall. | $2 \cdot \frac{precision \cdot recall}{precision+recall}$ | F1-score |
| How much more likely it is to achieve a positive test result in disease samples than in healthy control samples. | $\frac{sensitivity}{1-especificity}$ | $LR+$ |
| How much less likely it is to obtain a negative test result in a disease sample when compared a healthy control sample. | $\frac{1-sensitivity}{especificity}$ | $LR-$ |
| The ratio of the probability of a positive result in a disease sample to the that of a positive result in a healthy control sample. | $LR + \frac{LR+}{LR-}$ | DOR |
| The model's ability to discriminate between positive and negative examples, measuring the compensation between the TP rate and PF rate for different thresholds. | - | AUC |

precision and recall. Once again, the RF exhibited exceptional performance on this oversampled dataset, achieving the highest accuracy and F1 score. The RF also achieved high sensitivity and specificity, indicating its ability to identify both positive and negative cases effectively. Furthermore, the RF achieved high AUC and DOR, demonstrating its remarkable diagnostic performance. It displayed remarkable diagnostic performance with high AUC, DOR, LR+ values and low LR - values. These results indicate the superiority of the RF in accurately identifying positive cases while minimizing false-negatives.

In Table 9, LR emerged as the top performer on the RUS dataset. It achieved the highest accuracy and F1 score, indicating its strong overall performance. LR also demonstrated relatively high sensitivity, suggesting its effectiveness in correctly identifying positive and negative cases. Moreover, LR exhibited high AUC and DOR, further confirming its diagnostic capabilities. While other models, such as the RF and SVM, also showed competitive metrics, logistic regression proved to be the best-performing model in this analysis of the RUS dataset. However, when considering the AUC, DOR, LR+, and LR- metrics, LR displayed moderate discriminative power and diagnostic performance. This suggests that while LR performs reasonably well in diagnosing medical conditions, it may not provide the same level of accuracy and reliability as the RF on an oversampled dataset.

As shown in Table 10, the RF continued to outperform the other models based on the metrics. It demonstrated exceptional diagnostic performance, as reflected by the nearly perfect AUC and significantly high DOR. Additionally, the RF achieved a remarkably high LR+, indicating a substantial increase in the odds of a positive diagnosis while maintaining an extremely low LR-, implying a minimal likelihood of false-negatives. These findings highlight the outstanding diagnostic capabilities of the RF when using SMOTE.

Similarly, as shown in Table 11, the RF outperformed the other models, showing strong discriminative power and diagnostic performance, as evidenced by the high AUC and DOR values. Furthermore, the RF achieved a high LR+, indicating a considerable increase in the odds of a positive diagnosis while maintaining a low LR-, implying a low probability of false-negatives.

Table 12 shows that the accuracy values are competitive, allowing synthetic data to be generated considering the class imbalance. The ability of the ADASYN technique to assign large weights to data that are difficult to classify contributes to its high sensitivity. This suggests the potential of the method to identify cases of cognitive impairment, even if that means sacrificing specificity (this metric remains stable, indicating the ability to detect normal cases). Regarding LR, the values show the ability of the method to improve diagnostic certainty, while managing the likelihood of false positives and negatives. In turn, AUC has competitive values, which emphasize the ability to distinguish between cognitive impairment and normal cases efficiently. Likewise, DOR values improve the likelihood of diagnosis, suggesting an improvement in the accuracy of detection of cognitive impairment.

In Tables 7, 8, 9, 10, 11 and 12, we also considered the Cohen Kappa coefficient [37], as it is crucial to evaluate the agreement between the results generated by the model and those of a human expert. This coefficient measures the degree of confidence in the model's classification, considering the agreement. A high Cohen's Kappa value describe that the classification result is consistent with the human assessment, thereby enhancing confidence in the accuracy of the diagnostic model. In this research, the Cohen Kappa results enable us to interpret that the results obtained using the sampling techniques are dependable in aiding the diagnosis of cognitive impairment using the minimental test.

Table 13 describes that RF, LR, and SVM have better results in terms of sensitivity, that is, the probability of classifying data corresponding to MCI when it actually belongs to the disease. Additionally, Table 13 shows a substantial improvement using the sampling techniques, as opposed to what is obtained with the original dataset, considerably improving the generalization ability.

**TABLE 7. Original dataset versus model mean (Std).**

| | Accuracy | F1 | Sensitivity | Specificity | LR+ | LR- | AUC | DOR | Cohen's Kappa |
|---|---|---|---|---|---|---|---|---|---|
| SVM | 0.903(0.04) | 0.295(0.053) | 0.213(0.055) | 0.967(0.001) | 6.458(1.939) | 0.814(0.058) | 0.5(0.0) | 8.157(3.151) | 0.574(0.152) |
| RF | 0.935(0.038) | 0.484(0.137) | 0.4(0.15) | 0.967(0.001) | 11.996(4.43) | 0.621(0.155) | 0.654(0.132) | 25.397(27.916) | 0.715(0.176) |
| XGB | 0.921(0.036) | 0.447(0.136) | 0.38(0.138) | 0.955(0.016) | 9.444(4.546) | 0.649(0.145) | 0.632(0.127) | 16.836(11.698) | 0.709(0.187) |
| KNN | 0.915(0.045) | 0.369(0.139) | 0.29(0.159) | 0.967(0.001) | 8.808(5.067) | 0.734(0.164) | 0.578(0.152) | 16.426(20.785) | 0.554(0.111) |
| LR | 0.924(0.041) | 0.424(0.134) | 0.338(0.132) | 0.967(0.001) | 10.183(4.011) | 0.685(0.137) | 0.604(0.114) | 16.862(10.199) | 0.69(0.179) |
| MLP | 0.903(0.04) | 0.295(0.053) | 0.213(0.055) | 0.967(0.001) | 6.458(1.939) | 0.814(0.058) | 0.5(0.0) | 8.157(3.151) | 0.5(0.0) |
| NB | 0.887(0.063) | 0.486(0.158) | 0.538(0.196) | 0.893(0.05) | 6.453(4.909) | 0.519(0.223) | 0.74(0.187) | 19.455(25.314) | 0.741(0.121) |
| DT | 0.924(0.031) | 0.496(0.132) | 0.454(0.138) | 0.943(0.026) | 9.54(4.829) | 0.578(0.143) | 0.683(0.128) | 19.421(15.294) | 0.724(0.132) |

**TABLE 8. ROS dataset versus model mean (Std).**

| | Accuracy | F1 | Sensitivity | Specificity | LR+ | LR- | AUC | DOR | Cohen's Kappa |
|---|---|---|---|---|---|---|---|---|---|
| SVM | 0.955(0.027) | 0.928(0.024) | 0.967(0.003) | 0.882(0.051) | 11.31(8.373) | 0.038(0.004) | 0.955(0.027) | 306.097(233.059) | 0.955(0.024) |
| RF | 0.996(0.01) | 0.962(0.011) | 0.967(0.003) | 0.958(0.015) | 25.471(6.113) | 0.035(0.003) | 0.996(0.009) | 740.4(188.835) | 0.996(0.008) |
| XGB | 0.989(0.014) | 0.957(0.014) | 0.967(0.003) | 0.947(0.026) | 21.931(8.273) | 0.035(0.003) | 0.989(0.014) | 631.45(248.473) | 0.99(0.013) |
| KNN | 0.989(0.014) | 0.957(0.013) | 0.967(0.003) | 0.946(0.027) | 21.916(8.761) | 0.035(0.003) | 0.989(0.014) | 620.975(239.486) | 0.983(0.014) |
| LR | 0.978(0.025) | 0.947(0.023) | 0.967(0.003) | 0.926(0.043) | 17.204(8.472) | 0.036(0.004) | 0.978(0.023) | 492.763(261.385) | 0.98(0.016) |
| MLP | 0.903(0.042) | 0.873(0.043) | 0.862(0.077) | 0.89(0.045) | 10.387(7.646) | 0.155(0.084) | 0.903(0.042) | 117.414(172.237) | 0.91(0.05) |
| NB | 0.816(0.094) | 0.752(0.118) | 0.661(0.163) | 0.927(0.034) | 10.87(5.461) | 0.367(0.181) | 0.815(0.091) | 82.603(175.886) | 0.861(0.055) |
| DT | 0.988(0.018) | 0.955(0.019) | 0.967(0.003) | 0.944(0.029) | 21.511(8.328) | 0.036(0.004) | 0.988(0.016) | 626.5(263.69) | 0.984(0.018) |

**TABLE 9. RUS dataset versus model mean (Std).**

| | Accuracy | F1 | Sensitivity | Specificity | LR+ | LR- | AUC | DOR | Cohen's Kappa |
|---|---|---|---|---|---|---|---|---|---|
| SVM | 0.867(0.125) | 0.717(0.084) | 0.694(0.122) | 0.752(0.064) | 2.963(0.879) | 0.41(0.163) | 0.893(0.099) | 9.1(5.268) | 0.88(0.134) |
| RF | 0.833(0.149) | 0.693(0.093) | 0.662(0.139) | 0.752(0.064) | 2.842(0.95) | 0.454(0.192) | 0.869(0.11) | 8.292(5.424) | 0.89(0.119) |
| XGB | 0.717(0.159) | 0.609(0.099) | 0.563(0.141) | 0.7(0.131) | 2.157(0.899) | 0.651(0.255) | 0.726(0.159) | 4.536(3.918) | 0.744(0.183) |
| KNN | 0.808(0.121) | 0.672(0.08) | 0.627(0.112) | 0.752(0.064) | 2.679(0.776) | 0.5(0.159) | 0.835(0.093) | 6.458(4.026) | 0.889(0.121) |
| LR | 0.875(0.138) | 0.726(0.08) | 0.707(0.117) | 0.752(0.064) | 3.042(0.915) | 0.395(0.171) | 0.904(0.096) | 9.708(5.34) | 0.918(0.107) |
| MLP | 0.767(0.178) | 0.637(0.123) | 0.591(0.172) | 0.752(0.064) | 2.537(1.012) | 0.546(0.234) | 0.814(0.131) | 6.373(4.9) | 0.87(0.123) |
| NB | 0.85(0.117) | 0.734(0.076) | 0.773(0.061) | 0.645(0.125) | 2.468(0.892) | 0.363(0.103) | 0.865(0.105) | 8.033(5.012) | 0.772(0.137) |
| DT | 0.733(0.122) | 0.602(0.106) | 0.542(0.147) | 0.752(0.064) | 2.245(0.59) | 0.605(0.183) | 0.766(0.103) | 4.258(1.887) | 0.766(0.164) |

**TABLE 10. SMOTE sampled dataset versus model mean (Std).**

| | Accuracy | F1 | Sensitivity | Specificity | LR+ | LR- | AUC | DOR | Cohen's Kappa |
|---|---|---|---|---|---|---|---|---|---|
| SVM | 0.994(0.009) | 0.96(0.009) | 0.955(0.016) | 0.967(0.003) | 28.777(2.649) | 0.047(0.016) | 0.994(0.008) | 682.1(204.863) | 0.987(0.018) |
| RF | 0.991(0.012) | 0.958(0.011) | 0.952(0.02) | 0.965(0.007) | 27.921(3.912) | 0.05(0.021) | 0.992(0.011) | 646.65(231.574) | 0.989(0.015) |
| XGB | 0.982(0.018) | 0.949(0.016) | 0.935(0.032) | 0.965(0.008) | 27.52(4.288) | 0.067(0.033) | 0.983(0.017) | 500.265(227.892) | 0.974(0.02) |
| KNN | 0.989(0.018) | 0.955(0.019) | 0.946(0.034) | 0.967(0.003) | 28.522(2.799) | 0.056(0.035) | 0.989(0.018) | 654.1(247.804) | 0.995(0.008) |
| LR | 0.995(0.008) | 0.961(0.008) | 0.956(0.014) | 0.967(0.003) | 28.838(2.703) | 0.045(0.015) | 0.995(0.008) | 703.7(196.588) | 0.992(0.011) |
| MLP | 0.968(0.023) | 0.937(0.022) | 0.947(0.02) | 0.928(0.038) | 16.447(7.45) | 0.057(0.022) | 0.969(0.022) | 333.542(204.961) | 0.969(0.027) |
| NB | 0.953(0.04) | 0.921(0.042) | 0.922(0.033) | 0.923(0.05) | 16.138(8.404) | 0.085(0.039) | 0.953(0.038) | 270.781(253.183) | 0.958(0.025) |
| DT | 0.969(0.02) | 0.937(0.019) | 0.934(0.037) | 0.942(0.025) | 18.8(6.932) | 0.07(0.038) | 0.97(0.02) | 320.56(160.272) | 0.964(0.018) |

**TABLE 11. SMOTETomek sampled dataset versus model mean (Std).**

| | Accuracy | F1 | Sensitivity | Specificity | LR+ | LR- | AUC | DOR | Cohen's Kappa |
|---|---|---|---|---|---|---|---|---|---|
| SVM | 0.988(0.014) | 0.955(0.014) | 0.95(0.024) | 0.962(0.015) | 26.699(5.577) | 0.052(0.024) | 0.989(0.013) | 610.633(251.412) | 0.989(0.013) |
| RF | 0.983(0.015) | 0.95(0.015) | 0.938(0.027) | 0.963(0.01) | 26.654(4.939) | 0.064(0.028) | 0.983(0.015) | 510.162(248.091) | 0.987(0.013) |
| XGB | 0.977(0.02) | 0.944(0.021) | 0.934(0.035) | 0.955(0.018) | 23.426(6.976) | 0.069(0.037) | 0.977(0.02) | 443.046(247.043) | 0.97(0.016) |
| KNN | 0.988(0.016) | 0.955(0.015) | 0.947(0.028) | 0.965(0.007) | 27.57(3.92) | 0.055(0.029) | 0.989(0.015) | 603.48(232.626) | 0.995(0.008) |
| LR | 0.996(0.007) | 0.963(0.006) | 0.96(0.012) | 0.966(0.003) | 28.758(2.872) | 0.041(0.012) | 0.997(0.007) | 746.85(173.528) | 0.994(0.009) |
| MLP | 0.966(0.027) | 0.936(0.026) | 0.95(0.022) | 0.92(0.042) | 15.808(8.643) | 0.055(0.025) | 0.966(0.027) | 347.45(232.577) | 0.967(0.023) |
| NB | 0.943(0.045) | 0.911(0.045) | 0.906(0.059) | 0.92(0.047) | 15.857(9.007) | 0.104(0.067) | 0.942(0.045) | 221.518(187.62) | 0.943(0.045) |
| DT | 0.961(0.017) | 0.929(0.018) | 0.926(0.033) | 0.933(0.036) | 17.617(7.841) | 0.078(0.034) | 0.961(0.018) | 249.907(110.906) | 0.959(0.02) |

**TABLE 12. ADASYN sampled dataset versus model mean (Std).**

| | Accuracy | F1 | Sensitivity | Specificity | LR+ | LR- | AUC | DOR | Cohen's Kappa |
|---|---|---|---|---|---|---|---|---|---|
| SVM | 0.995(0.008) | 0.961(0.007) | 0.956(0.013) | 0.967(0.004) | 29.167(3.262) | 0.045(0.014) | 0.995(0.008) | 700.05(193.873) | 0.988(0.014) |
| RF | 0.99(0.012) | 0.956(0.011) | 0.951(0.018) | 0.964(0.009) | 27.457(5.02) | 0.051(0.019) | 0.991(0.012) | 618.117(235.454) | 0.989(0.012) |
| XGB | 0.973(0.018) | 0.941(0.017) | 0.943(0.022) | 0.94(0.025) | 18.665(7.64) | 0.061(0.024) | 0.973(0.019) | 376.246(250.752) | 0.98(0.015) |
| KNN | 0.987(0.015) | 0.952(0.014) | 0.941(0.026) | 0.967(0.004) | 28.687(3.168) | 0.061(0.027) | 0.987(0.014) | 569.492(245.435) | 0.991(0.016) |
| LR | 0.995(0.008) | 0.961(0.007) | 0.956(0.013) | 0.967(0.004) | 29.167(3.262) | 0.045(0.014) | 0.995(0.008) | 700.05(193.873) | 0.994(0.009) |
| MLP | 0.969(0.024) | 0.936(0.025) | 0.951(0.016) | 0.926(0.044) | 17.576(8.706) | 0.053(0.017) | 0.97(0.021) | 358.505(220.189) | 0.961(0.028) |
| NB | 0.951(0.037) | 0.919(0.039) | 0.931(0.038) | 0.909(0.05) | 15.06(9.566) | 0.077(0.044) | 0.95(0.037) | 284.221(258.24) | 0.957(0.026) |
| DT | 0.967(0.021) | 0.935(0.02) | 0.939(0.026) | 0.932(0.037) | 18.322(9.227) | 0.065(0.027) | 0.967(0.021) | 326.928(228.209) | 0.964(0.03) |

**TABLE 13. Analysis of techniques versus sampling.**

| Technique | Sampling | Accuracy | Sensitivity |
|---|---|---|---|
| RF | Original Data | 0.935 | 0.400 |
| RF | ROS | 0.996 | 0.967 |
| NB | RUS | 0.85 | 0.773 |
| LR | SMOTE | 0.995 | 0.956 |
| LR | SMOTETomek | 0.996 | 0.960 |
| SVM/LR | ADASYN | 0.995 | 0.956 |

## IV. DISCUSSION

The RUS technique reduces the overall accuracy. While the sensitivity values improve, the decrease in the accuracy values indicates that the ability of the model to classify balanced and unbalanced cases is impaired. This implies that the model can identify cases of cognitive impairment but may misclassify some normal cases. The sensitivity values indicate that the model can correctly classify cases of cognitive impairment. This is significant because it is possible to address the problem of data imbalance, which reduces the overrepresentation of the majority class. Regarding specificity, the values of this metric are low, as the sensitivity values are high. This implies that while the model becomes good at identifying positive cases, there is a correspondence in increasing false-positives among normal cases. Sensitivity and specificity influence the LR+ and LR- ratios. A high LR+ and a low LR- mean that the model has improved its diagnostic ability for cognitive impairment, but the false-positive likelihood increases. The AUC metric indicates that the reduction in accuracy is reflected by a low AUC, demonstrating that the discriminative ability of the model is lowered. Finally, the trade-off between sensitivity and specificity implies that the DOR metric increases. However, the risk of false-positives also increases.

For the ROS technique, the good results are achieved with respect to the accuracy metric. However, these results are accompanied by a decrease in specificity, indicating that the overall classification is improved, but there is an increase in false-positives. The sensitivity is stable, which implies that the model focuses on balancing the representation of the classes. However, the reduction in specificity indicates that the model continues to detect normal cases as cases of cognitive impairment. The trade-off between sensitivity and specificity influences LR+ and LR-, indicating a trade-off between high true-positive and false-positive rates. In turn, the improvement in accuracy leads to a high AUC, which improves the discriminative ability of the model. Finally, the increase in accuracy results in an increase in the DOR, which improves the diagnostic accuracy.

SMOTE is highly accurate because its ability to generate synthetic data improves the balance between classes, resulting in accurate predictions for both balanced and unbalanced cases. Good performance is achieved in terms of the sensitivity metric, indicating that this method is appropriate for identifying cases of cognitive impairment. This aligns with the objective of addressing class imbalance and improving the detection of positive cases. The specificity metric remains stable, indicating that SMOTE also classifies the normal cases. In turn, the balance between sensitivity and specificity allows for adequate values for LR+ and LR-. In this sense, SMOTE contributes to better diagnosis while mitigating the risk of false-positives and false-negatives. Regarding the AUC metric, the results of this metric indicate that SMOTE is effective in distinguishing between cognitive impairment cases and normal cases. Regarding the DOR metric, the values obtained are relevant, indicating SMOTE's ability to improve the diagnostic accuracy across the entire dataset.

With regard to SMOTETomek, the accuracy results are robust, implying that the technique is successful in refining the dataset by removing Tomek links and improving the separability of the data. The results for the sensitivity metric suggest that the model improves its ability to identify cases of cognitive impairment. In relation to sensitivity, the values of the metric indicate that SMOTETomek preserves the model's ability to identify unbalanced classes. In relation to LR+ and LR-, an improved diagnosis is achieved without compromising this balance. The AUC values are good, which also demonstrates the discriminative ability of this technique. Finally, the DOR metric demonstrates that SMOTE-Tomek can improve the diagnostic accuracy and reliability.

The accuracy results obtained using the ADASYN technique indicate that the obtained values are competitive, as they describe the ability to generate synthetic data dealing with class imbalance. The ability of ADASYN to assign large weights to hard-to-classify data contributes to high values of the sensitivity metric. This suggests the potential of the method to identify cases of cognitive impairment, even if that means sacrificing specificity, as this metric describes stable results, indicating the ability to detect normal cases. Regarding LR+ and LR−, the values show the ability of the method to improve the diagnostic certainty while managing the likelihood of false-positives and false-negatives. In turn, the values obtained for the AUC metric indicate that ADASYN can efficiently distinguish between cognitive impairment and normal cases. Finally, the values obtained for the DOR metric suggest an improvement in the accuracy of cognitive impairment detection.

Concerning performance, LR and SVM demonstrate strong performance in diagnostic metrics, particularly in discerning patterns indicative of the target class, in this instance, MCI. This success in applying probability-based techniques to distinguish the disease effectively increase confidence in their utility for diagnostic tasks, particularly in situations where data is difficult to differentiate or distinctions are minimal.

### A. PRINCIPAL RESULTS

After the evaluation of various resampling techniques within the context of mild cognitive impairment diagnosis using an unbalanced MMSE dataset, it is observed that these techniques perform well for classification techniques. Among

the techniques explored, SMOTE performs best, achieving superior classification performance for the metrics used, including accuracy, precision, recall, F1, and AUC. Its ability to generate synthetic samples to balance the dataset while preserving the underlying distribution of data contributes to its excellent performance. In this regard, the RF shows a better fit for unbalanced datasets. This method corresponds to an ensemble that uses multiple decision trees to capture patterns in the data and obtain a robust prediction. These findings suggest that SMOTE and the RF can significantly improve the accuracy and reliability of cognitive impairment diagnosis. These results reflect the importance of dealing with data imbalance and using advanced classification algorithms that improve the predictive ability of learning models.

From a clinical perspective, this study offers important results in the application of machine learning techniques for the diagnosis of cognitive impairment. In particular, SMOTE using the RF has the potential to significantly improve diagnostic outcomes, helping health care professionals make more informed and timely decisions regarding patient care. As medicine continues to be aided by technological solutions, our study will play a pivotal role in approaches to cognitive impairment detection and patient support.

### B. LIMITATIONS

This study is limited by the relatively small size of our dataset. This affects the generalizability of the methods and may not represent the diversity and complexity of cognitive impairment in older adults. On the other hand, we use data with some level of bias due to selection criteria factors or patient location. Regarding class imbalance, despite applying some balancing techniques, the data may still have some level of imbalance, which affects the effectiveness of the methods. To mitigate these issues, we followed the recommendation of Button et al. [38] regarding describing methods and findings in a transparent manner. In this regard, the machine learning techniques used in our study are implemented in highly cited programming libraries used in data research (such as scikit-learn Python library). Additionally, the findings we reported are supported on the basis of the data described in the article to foster reproducibility of the results.

## V. RELATED WORK

Belarouci and Chikh [39] discussed how imbalanced data are related to medical diagnosis. The authors mentioned that because of the distribution of the imbalanced data, it is difficult to obtain good performance for most of them using traditional classifiers where a balanced distribution exists. In addition, Sribhashyam et al. [40] analysed complex imbalanced data with high dimensionality and big data. In contrast, the problem of imbalanced data can be combined with incomplete data in a dataset named Western Medicine and Symptom Prediction using a multi-instance neural network architecture. There are problems associated with computer vision using machine learning with imbalanced data.

Lee et al. [41] evaluated different sampling techniques using the naive Bayes method. Similarly, He et al. [42] noticed that the hyperplane is biased to the majority class for an SVM using imbalanced data, leading to more false-negative predictions. Feltes et al. [18] used other classifiers against imbalanced data for microarray gene expression analysis. Kadir et al. [43] discussed how k-nearest neighbour is biased towards the majority class in the training set, and Cieslak et al. [44] discussed decision trees. Brown and Mues [45] and Chang and Chawla [46] analysed how gradient boosting outperforms SVM, decision trees, and kNN for problems with imbalanced data.

Regarding cognitive assessment and the MMSE, Flaxman and Vos [47] mentioned that developing machine learning approaches for this task requires considerable experience and clinical correlation. Machine learning techniques have constraints such as fairness, accountability, transparency, privacy, explainability, and causal inference. Almubark et al. [48] used a multilayer perceptron (MLP) to compare the classification performance using behavioural data. Youn et al. [49] used the Korean Dementia Screening Questionnaire (KDSQ) and the MMSE, utilizing 24 variables, including education, sex, age, and hypertension. So et al. [50] used naive Bayes, bagging, and demographic data to predict normal, MCI, and dementia in a quick, inexpensive, and reliable way to detect dementia in its early stages and to increase accuracy.

Jun et al. [3] develop a multiple linear regression model to predict mild cognitive impairment using a combination of bioimpedance variables and the Korean Mini-Mental State Examination total score. The authors compared the accuracy of the model with SNSB-II (Seoul Neuropsychological Screening Battery) domain scores using the area under the receiver operating characteristic. In addition, they analyzed the performance of the model using several machine learning models. Similarly, García-Gutiérrez et al. [51] implement several machine learning models to identify individuals without cognitive impairment (subjective cognitive impairment), with mild cognitive impairment, and dementia due to Alzheimer's disease. The authors described models that are capable of predicting performance in cognitive domains.

## VI. CONCLUSION

This study evaluated machine learning techniques on unbalanced data related to MMSE test responses in elderly patients. The techniques used were naive Bayes, support vector machine, k-nearest neighbour, decision trees, random forest, logistic regression, and multilayer perceptron. The dataset consists of data from the EHomeSenior project, the main objective of which is to monitor older adults. Because the dataset was unbalanced, we applied balancing techniques to avoid bias in our study, namely, RUS, ROS, SMOTE, SMOTETomek, and ADASYN. To evaluate the performance of the machine learning techniques, we used seven metrics: sensitivity, specificity, F1-score, LR+, LR−, DOR, and

AUC. The results obtained indicate that SMOTE together with the random forest significantly improves the accuracy of diagnosing cognitive impairment.

The primary advantage of this research lies in the application of sampling methods to datasets that are inherently unbalanced. It is widely acknowledged that obtaining disease data, particularly for complex conditions where distinguishing one patient from another is challenging, is difficult. Consequently, a significant portion of such data is typically derived from healthy individuals.

The results obtained provide the opportunity to use techniques in suspected patients according to the results of the MMSE test and provide a new tool to accept results or discard them. Independent of the sampling technique, they decrease the bias for the majority class and improve the general classification; however, no single method alone can achieve the best performance. The data must be evaluated for every case. More importantly, every case must be analysed with more than one metric, with no trust in only one of them. This is especially important when analysing clinical data and depends on the classifier and sampling method.

## REFERENCES


[1] N. García-Casares, P. Gallego Fuentes, M. Á. Barbancho, R. López-Gigosos, A. García-Rodríguez, and M. Gutiérrez-Bedmar, "Alzheimer's disease, mild cognitive impairment and Mediterranean diet. A systematic review and dose-response meta-analysis," *J. Clin. Med.*, vol. 10, no. 20, p. 4642, Oct. 2021.

[2] P. S. Aisen, J. Cummings, C. R. Jack, J. C. Morris, R. Sperling, L. Frölich, R. W. Jones, S. A. Dowsett, B. R. Matthews, J. Raskin, P. Scheltens, and B. Dubois, "On the path to 2025: Understanding the Alzheimer's disease continuum," *Alzheimer's Res. Therapy*, vol. 9, no. 1, pp. 1–10, Dec. 2017.

[3] M.-H. Jun, B. Ku, K. Kim, K. H. Lee, and J. U. Kim, "A screening method for mild cognitive impairment in elderly individuals combining bioimpedance and MMSE," *Frontiers Aging Neurosci.*, vol. 16, Jan. 2024, Art. no. 1307204.

[4] R. C. Petersen, J. C. Stevens, M. Ganguli, E. G. Tangalos, J. L. Cummings, and S. T. DeKosky, "Practice parameter: Early detection of dementia: Mild cognitive impairment (an evidence-based review) [RETIRED]: Report of the quality standards subcommittee of the American academy of neurology," *Neurology*, vol. 56, no. 9, pp. 1133–1142, May 2001.

[5] S. Farzana and N. Parde, "Exploring MMSE score prediction using verbal and non-verbal cues," in *Proc. Interspeech*, Oct. 2020, pp. 2207–2211.

[6] I. Arevalo-Rodriguez, N. Smailagic, M. Roqué i Figuls, A. Ciapponi, E. Sanchez-Perez, A. Giannakou, O. L. Pedraza, X. B. Cosp, and S. Cullum, "Mini-mental state examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI)," *BJPsych Adv.*, vol. 21, no. 6, p. 362, Nov. 2015.

[7] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, "A survey on addressing high-class imbalance in big data," *J. Big Data*, vol. 5, no. 1, pp. 1–30, Dec. 2018.

[8] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, pp. 1–54, Dec. 2019.

[9] S. Ayoub, Y. Gulzar, J. Rustamov, A. Jabbari, F. A. Reegu, and S. Turaev, "Adversarial approaches to tackle imbalanced data in machine learning," *Sustainability*, vol. 15, no. 9, p. 7097, Apr. 2023.

[10] B. C. Feltes, E. B. Chandelier, B. I. Grisci, and M. Dorn, "CuMiDa: An extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research," *J. Comput. Biol.*, vol. 26, no. 4, pp. 376–386, Apr. 2019.

[11] B. C. Feltes, J. D. F. Poloni, I. J. G. Nunes, S. S. Faria, and M. Dorn, "Multi-approach bioinformatics analysis of curated omics data provides a gene expression panorama for multiple cancer types," *Frontiers Genet.*, vol. 11, Nov. 2020, Art. no. 586602.

[12] C. Taramasco, T. Rodenas, F. Martinez, P. Fuentes, R. Munoz, R. Olivares, V. H. C. De Albuquerque, and J. Demongeot, "A novel monitoring system for fall detection in older people," *IEEE Access*, vol. 6, pp. 43563–43574, 2018.

[13] G. Márquez, A. Veloz, J.-G. Minonzio, C. Reyes, E. Calvo, and C. Taramasco, "Using low-resolution non-invasive infrared sensors to classify activities and falls in older adults," *Sensors*, vol. 22, no. 6, p. 2321, Mar. 2022.

[14] C. Rimassa and C. Taramasco, "When age becomes a plus: Seniors and cognitive reserve," *Int. Phys. Med. Rehabil. J.*, vol. 6, no. 3, pp. 70–72, Nov. 2021.

[15] N. Hounsome, M. Orrell, and R. T. Edwards, "EQ-5D as a quality of life measure in people with dementia and their carers: Evidence and key issues," *Value Health*, vol. 14, no. 2, pp. 390–399, Mar. 2011.

[16] M. F. Folstein, S. E. Folstein, and P. R. McHugh, "'Mini-mental state': A practical method for grading the cognitive state of patients for the clinician," *J. Psychiatric Res.*, vol. 12, no. 3, pp. 189–198, 1975.

[17] A. J. Mitchell, "The mini-mental state examination (MMSE): Update on its diagnostic accuracy and clinical utility for cognitive disorders," *Cognit. Screening Instruments: A Practical Approach*, pp. 37–48, 2017. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-44775-9_3#citeas

[18] B. C. Feltes, J. D. F. Poloni, and M. Dorn, "Benchmarking and testing machine learning approaches with BARRA: CuRDa, a curated RNA-seq database for cancer research," *J. Comput. Biol.*, vol. 28, no. 9, pp. 931–944, Sep. 2021.

[19] B. I. Grisci, B. C. Feltes, and M. Dorn, "Neuroevolution as a tool for microarray gene expression pattern identification in cancer research," *J. Biomed. Informat.*, vol. 89, pp. 122–133, Jan. 2019.

[20] S. Inampudi, G. Johnson, J. Jhaveri, S. Niranjan, K. Chaurasia, and M. Dixit, "Machine learning based prediction of H1N1 and seasonal flu vaccination," in *Proc. Adv. Computing, 10th Int. Conf.*, 2020, pp. 139–150.

[21] P. Ormeño, G. Márquez, C. Guerrero-Nancuante, and C. Taramasco, "Detection of COVID-19 patients using machine learning techniques: A nationwide Chilean study," *Int. J. Environ. Res. Public Health*, vol. 19, no. 13, p. 8058, Jun. 2022.

[22] Y. Huang and L. Li, "Naive Bayes classification algorithm based on small sample set," in *Proc. IEEE Int. Conf. Cloud Comput. Intell. Syst.*, Sep. 2011, pp. 34–39.

[23] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[24] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: A review of classification and combining techniques," *Artif. Intell. Rev.*, vol. 26, no. 3, pp. 159–190, Nov. 2006.

[25] P. Harrington, *Machine Learning in Action*. New York, NY, USA: Simon and Schuster, 2012.

[26] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," *Ensemble Mach. Learn., Methods Appl.*, pp. 157–175, 2012. [Online]. Available: https://link.springer.com/chapter/10.1007/978-1-4419-9326-7_5#citeas

[27] A. Géron, *Hands-on Machine Learning With Scikit-Learn, Keras, and TensorFlow*. Sebastopol, CA, USA: O'Reilly Media, Inc., 2022.

[28] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[29] Y. Başţanlar and M. Özuysal, "Introduction to machine learning," in *MiRNomics: MicroRNA Biology*. Cham, Switzerland: Springer, 2014, pp. 105–128.

[30] N. Tomašev and D. Mladenić, "Class imbalance and the curse of minority hubs," *Knowl.-Based Syst.*, vol. 53, pp. 157–172, Nov. 2013.

[31] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, May 2017.

[32] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[33] M. Ringnér, "What is principal component analysis?" *Nature Biotechnol.*, vol. 26, no. 3, pp. 303–304, Mar. 2008.

[34] A. Tharwat, "Classification assessment methods," *Appl. Comput. Informat.*, vol. 17, no. 1, pp. 168–192, Jan. 2021.

[35] A.-M. Šimundić, "Measures of diagnostic accuracy: Basic definitions," *Electron. J. Int. Fed. Clin. Chem. Lab. Med.*, vol. 19, no. 4, p. 203, 2009.

[36] A. S. Glas, J. G. Lijmer, M. H. Prins, G. J. Bonsel, and P. M. M. Bossuyt, "The diagnostic odds ratio: A single indicator of test performance," *J. Clin. Epidemiology*, vol. 56, no. 11, pp. 1129–1135, Nov. 2003.
[37] S. M. Vieira, U. Kaymak, and J. M. Sousa, "Cohen's Kappa coefficient as a performance measure for feature selection," in *Proc. Int. Conf. Fuzzy Syst.*, 2010, pp. 1–8.
[38] K. S. Button, J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafò, "Power failure: Why small sample size undermines the reliability of neuroscience," *Nature Rev. Neurosci.*, vol. 14, no. 5, pp. 365–376, May 2013.
[39] S. Belarouci and M. A. Chikh, "Medical imbalanced data classification," *Adv. Sci., Technol. Eng. Syst. J.*, vol. 2, no. 3, pp. 116–124, 2017.
[40] S. Sribhashyam, S. Koganti, M. V. Vineela, and G. Kalyani, "Medical diagnosis for incomplete and imbalanced data," in *Proc. Intelligent Data Engineering and Analytics*. Cham, Switzerland: Springer, 2021, pp. 491–499.
[41] M. S. Lee, J.-K. Rhee, B.-H. Kim, and B.-T. Zhang, "Aesnb: Active example selection with naive Bayes classifier for learning from imbalanced biomedical data," in *Proc. 9th IEEE Int. Conf. Bioinf. Bioeng.*, 2009, pp. 15–21.
[42] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Jun. 2008, pp. 1322–1328.
[43] M. E. Kadir, P. S. Akash, S. Sharmin, A. A. Ali, and M. Shoyaib, "A proximity weighted evidential k nearest neighbor classifier for imbalanced data," in *Proc. 24th Pacific–Asia Conf.*, 2020, pp. 71–83.
[44] D. A. Cieslak and N. V. Chawla, "Learning decision trees for unbalanced data," in *Proc. Mach. Learn. Knowl. Discovery Databases, Eur. Conf.*, 2008, pp. 241–256.
[45] I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Expert Syst. Appl.*, vol. 39, no. 3, pp. 3446–3453, Feb. 2012.
[46] Y.-C. Chang, K.-H. Chang, and G.-J. Wu, "Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions," *Appl. Soft Comput.*, vol. 73, pp. 914–920, Dec. 2018.
[47] A. D. Flaxman and T. Vos, "Machine learning in population health: Opportunities and threats," *PLOS Med.*, vol. 15, no. 11, Nov. 2018, Art. no. e1002702.
[48] I. Almubark, L.-C. Chang, K. F. Shattuck, T. Nguyen, R. S. Turner, and X. Jiang, "A 5-min cognitive task with deep learning accurately detects early Alzheimer's disease," *Frontiers Aging Neurosci.*, vol. 12, Dec. 2020, Art. no. 603179.
[49] Y. C. Youn, S. H. Choi, H.-W. Shin, K. W. Kim, J.-W. Jang, J. J. Jung, G.-Y.-R. Hsiung, and S. Kim, "Detection of cognitive impairment using a machine-learning algorithm," *Neuropsychiatric Disease Treat.*, vol. 14, pp. 2939–2945, Nov. 2018.
[50] A. So, D. Hooshyar, K. Park, and H. Lim, "Early diagnosis of dementia from clinical data by machine learning techniques," *Appl. Sci.*, vol. 7, no. 7, p. 651, Jun. 2017.
[51] F. García-Gutiérrez et al., "Unveiling the sound of the cognitive status: Machine learning-based speech analysis in the Alzheimer's disease spectrum," *Alzheimer's Res. Therapy*, vol. 16, no. 1, pp. 1–20, Feb. 2024.

**PABLO ORMEÑO** received the Ph.D. degree in computer science and informatics engineering from Federico Santa María Technical University, in 2021. Since 2019, he has been actively engaged in research with Viña del Mar University and has contributed to various projects in the fields of computer science and artificial intelligence. His work is focused on advancing the understanding and application of machine-learning techniques to solve complex problems, with a particular emphasis on natural language processing. His research interests include machine learning, deep learning, and natural language processing.

**GASTÓN MÁRQUEZ** is currently pursuing the Ph.D. degree in informatics engineering with Federico Santa María Technical University, Chile. He is a Professor with Universidad del Bío-Bío, Chile. Before becoming the Ph.D. student, he was with financial companies for five years. He has published in several international conferences and has participated in international software architecture schools. He participated as a Research Visitor with Rochester Institute of Technology (RIT), Rochester, NY, USA, and Université de Technologie de Compiégne (UTC), Compiégne, France. His research interests include architectural tactics, patterns, microservice architectures, technical debt, and security in telehealth systems.

**CARLA TARAMASCO** received the B.Eng. degree in computer engineering from Universidad de Valparaíso, Chile, in 2001, the M.Sc. degree in cognitive science from École Normale Supérieure, in 2006, and the Ph.D. degree (summa cum laude) from École Polytechnique, France, in 2011. She is currently a Full Professor with the Faculty of Engineering, Andrés Bello University, and the Director of the Institute of Technology for Innovation in Health and Wellbeing (ITiSB). Her research interests include health and complex social systems. She is also a member of the Executive Committee, National Centre for Health Information Systems.