Computer Methods and Programs in Biomedicine Update 5 (2024) 100132

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine Update

journal homepage: www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine-update

Machine learning from real data: A mental health registry case study

Elisabetta Gentili [a], Giorgia Franchini [b,*], Riccardo Zese [c], Marco Alberti [d], Maria Ferrara [e,f,g], Ilaria Domenicano [e], Luigi Grassi [e,f]

[a] *Department of Engineering, University of Ferrara, Ferrara, Italy*
[b] *Department of Physics, Informatics and Mathematics, University of Modena and Reggio Emilia, Modena, Italy*
[c] *Department of Chemical, Pharmaceutical and Agricultural Sciences, University of Ferrara, Ferrara, Italy*
[d] *Department of Mathematics and Computer Science, University of Ferrara, Ferrara, Italy*
[e] *Institute of Psychiatry, Department of Neuroscience and Rehabilitation, University of Ferrara, Ferrara, Italy*
[f] *Integrated Department of Mental Health and Pathological Addictions, Ferrara Local Health Trust, Ferrara, Italy*
[g] *Department of Psychiatry, Yale School of Medicine, New Haven, USA*

ARTICLE INFO

Keywords:
Healthcare
Machine learning
Imbalanced dataset
Mental health

ABSTRACT

Imbalanced datasets can impair the learning performance of many Machine Learning techniques. Nevertheless, many real-world datasets, especially in the healthcare field, are inherently imbalanced. For instance, in the medical domain, the classes representing a specific disease are typically the minority of the total cases. This challenge justifies the substantial research effort spent in the past decades to tackle data imbalance at the data and algorithm levels. In this paper, we describe the strategies we used to deal with an imbalanced classification task on data extracted from a database generated from the Electronic Health Records of the Mental Health Service of the Ferrara Province, Italy. In particular, we applied balancing techniques to the original data, such as random undersampling and oversampling, and Synthetic Minority Oversampling Technique for Nominal and Continuous (SMOTE-NC). In order to assess the effectiveness of the balancing techniques on the classification task at hand, we applied different Machine Learning algorithms. We employed cost-sensitive learning as well and compared its results with those of the balancing methods. Furthermore, a feature selection analysis was conducted to investigate the relevance of each feature. Results show that balancing can help find the best setting to accomplish classification tasks. Since real-world imbalanced datasets are increasingly becoming the core of scientific research, further studies are needed to improve already existing techniques.

1. Introduction

In classification problems, learning from imbalanced data, where one class is under-represented, poses peculiar challenges to Machine Learning (ML) algorithms and has received considerable attention from the community [1,2]. The degree of imbalance is considered mild when the proportion of the minority class is 20%–40% of the dataset, moderate when the proportion is 1%–20%, and extreme when the proportion is <1%. One of the most commonly used measures to describe the imbalance of a dataset is the Imbalance Ratio (IR), defined as the ratio between the number of examples in the majority class and the number of examples in the minority class. When IR=1, the dataset is perfectly balanced, while an IR> 1 indicates that the dataset is imbalanced, and the higher the IR, the greater the imbalance.

Imbalanced datasets are widespread and also expected in those fields where the class of interest is usually underrepresented, such as fraud detection [3], or intrusion detection [4,5], where the number of fraudulent situations is usually extremely smaller than legal ones. Class imbalance is also a major issue in healthcare, especially in diagnosis prediction models where individuals affected by the disease of interest are often the minority of the sample [6–8].

Class imbalance heavily affects the performances of standard supervised ML techniques [9] used for classification tasks, such as decision trees and Support Vector Machines (SVMs), as they often assume that classes are equally distributed. Consider, for example, the following example: given an imbalanced dataset with 90% of negative examples, a ML model that classifies every example as negative will achieve an accuracy of 90%, because the accuracy does not take into account data imbalance. Since using accuracy alone as a performance metric might result in misleading conclusions, other metrics are computed as well, such as balanced accuracy, recall, and F1-score [9,10].

[*] Corresponding author.
E-mail address: giorgia.franchini@unimore.it (G. Franchini).

https://doi.org/10.1016/j.cmpbup.2023.100132
Received 13 August 2023; Received in revised form 18 December 2023; Accepted 24 December 2023
Available online 30 December 2023

The problem of imbalanced classes can be solved using data-level and algorithm-level methods. Data-level methods, such as random resampling or the Synthetic Minority Oversampling Technique (SMOTE) [11], deal with the problem by modifying the number of training samples in order to decrease the imbalance ratio. Random undersampling discards samples from the majority group, which might result in the loss of valuable information. On the other hand, random oversampling duplicates samples from the minority class, introducing, however, a possibility of overfitting. SMOTE is a type of oversampling that creates new examples close to real ones by randomly selecting an example of the minority class, computing its k-nearest neighbors, and then randomly choosing one of them to add to the dataset. For this reason, it should perform better than standard random undersampling and oversampling. SMOTE for Nominal and Continuous (SMOTE-NC) is a particular version of SMOTE in which the dataset to resample contains both numerical and categorical features. In all the cases detailed above, the goal is to use one or more techniques in such a way that the smallest amount of data is lost and the duplicated data are useful to strengthen class boundaries and improve discrimination. Once these techniques are applied, the analysis will turn to the robustness of the ML methodologies.

On the other hand, unlike data-level methods, algorithm-level methods such as cost-sensitive learning [12] do not alter the training dataset distribution, instead, they change the learning rule, in order to keep track of the imbalance ratio. Penalties are given to the different classes in order to increase the cost of misclassification for members of the minority group (i.e., increasing its importance) and decrease the cost of misclassification for members of the majority group (i.e., decreasing its importance). The costs corresponding to false positive and false negative errors are adjusted in the same way.

Starting from a real-world mental health records dataset, we will analyze how different data balancing strategies impacted on the performance of different supervised ML algorithms. This manuscript is organized as follows. Section 2 discusses related work. Section 3 describes the dataset preprocessing and the methodologies used to deal with the dataset. Section 4 describes the experimental results. Finally, Section 5 discusses the findings and Section 6 provides the conclusions of the study.

2. Related work

In this section, we will try to include some of the approaches already known to the field to address dataset imbalance in the medical context and to highlight their differences and similarities with respect to our context and approach.

Rahman and Davis [13] analyzed data from a cardiovascular context, in particular they examined a dataset that had a balancing percentage of about 17%, quite similar to the case examined in this paper. The balancing techniques they applied are also similar to those analyzed in this paper, but they considered a set containing only 823 patients, far less than our case study (see Section 3).

Among the published literature on this challenge, Belarouci and Chikh’s study stood out for several reasons [14]. The authors proposed a new complex technique to deal with imbalanced datasets and apply it to different ML and Deep Learning (DL) techniques. On the other hand, they only considered datasets containing less than 1,000 entities. Conversely, in this work, we applied the same ML techniques, but taking into account a new dataset and real data, not prepared *ad hoc* for classification. Similarly, Li et al. [15] presented a work that dealt with the approach towards imbalanced datasets in the medical field, considering in this case milder unbalancing percentages for datasets with cardinality of a few hundreds.

Gerych et al. [8] tried to classify depressed vs not-depressed users. The authors stated that ‘Machine learning classification on significantly imbalanced datasets can be reformulated as an anomaly detection problem, where instances of the minority class are considered anomalies’. They also proposed a DL technique called autoencoder to address the problem of classification on imbalanced datasets, however, they used only datasets available in the literature.

In their work, Khushi et al. [16] applied different balancing methods to state-of-the-art ML algorithms (logistic regression, random forest, and LinearSVC) on different imbalanced medical datasets, and showed that class imbalance learning can effectively improve the classification ability of the model. Zeng et al. [17] combined SMOTE [11] with Tomek links technique [18] to preprocess imbalanced medical data. They compared the performance of classifiers with this combination to the classifiers who applied only SMOTE on three imbalanced datasets of different diseases. The results obtained using both SMOTE and Tomek links techniques were superior in terms of all metrics considered by the authors up to 4 percentage points compared to those obtained using only SMOTE.

Regarding algorithm-level methods and cost-sensitive learning, Sheng et al. [19] compared different algorithms for cost-sensitive learning applied to several heterogeneous domains. Focusing on healthcare, different works studied the application of cost-sensitive learning techniques to imbalanced medical data [20–23]. In particular, in [20] the authors applied cost-sensitive learning to four widespread ML algorithms, namely logistic regression, decision tree, extreme gradient boosting (XGBoost), and random forest, They compared the results with the performances of the standard version of the algorithms. They tested the algorithms on four popular medical datasets and found that the cost-sensitive methods performed better compared to the standard algorithms. The authors of [22] proposed a cost-sensitive XGBoost model and applied it to four breast cancer imbalanced datasets. In [21] the authors proposed a class weights voting based on a random forest and obtained good results on five different imbalanced medical datasets. Finally, Ali et al. [23] developed a cost-sensitive ensemble feature ranking. Each of these works showed that applying cost-sensitive learning improved the performances of the chosen ML algorithms used on imbalanced medical data.

3. Methods

3.1. Data source

The data used in this study come from FEPSY (FErrara-PSYchiatry) [24], a research database created by extracting information from EFESO, the Electronic Health Record (EHR) employed by the Local Health Trust of Ferrara for Mental Health in Adults, which covers a catchment of 342,061 inhabitants. The Local Health Trust of Ferrara began to accurately collect data regarding mental health services in the late 1970s, with the constitution of the Departments of Mental Health in the Region in agreement with the Italian mental health reform (13 May 1978, Law 180) [25], and introduced the first EHR in 1991, as the availability of informatics had expanded by then. In the following years, many different EHRs were adopted, and each new EHR replaced the previous one by importing already existing data and adding new features to the software. FEPSY includes 46,222 individuals who had access to the mental health services of the province of Ferrara from 1991 to February 2021. Included data are both socio-demographic and clinical (e.g., medical records, diagnoses, medical services, treatment plans, psychometric tests, medication prescriptions, and distribution).

In this study, we tried to classify patients with a diagnosis of psychotic spectrum disorders (295.xx, 297.xx, and 298.xx – excl. 298.0 – International Classification of Diseases, Ninth Revision (ICD-9) codes [26]), in order to find potential predictors for these illnesses. For this study, we selected a subset of the subjects included in FEPSY. First, we excluded those individuals who had at least one medical record or product (e.g., consulting, hospitalization) with inconsistent start and end dates (i.e., it ended before it started). After that, we excluded those individuals who had not received any diagnosis. Then we excluded those who were still receiving care as of February 2021. Finally, we

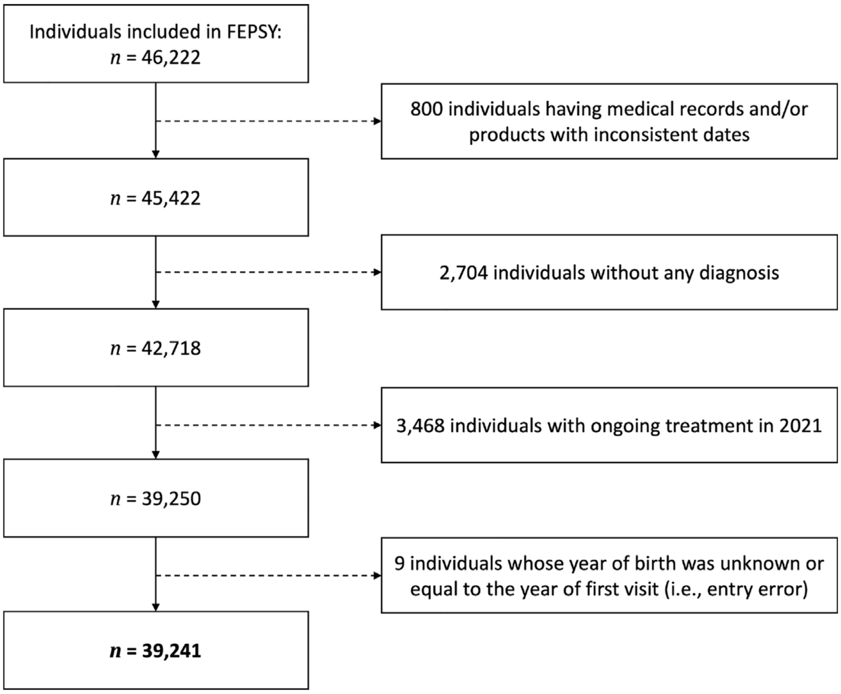

Individuals included in FEPSY:
n = 46,222
800 individuals having medical records and/or products with inconsistent dates
n = 45,422
2,704 individuals without any diagnosis
n = 42,718
3,468 individuals with ongoing treatment in 2021
n = 39,250
9 individuals whose year of birth was unknown or equal to the year of first visit (i.e., entry error)
n = 39,241

Fig. 1. Graphic representation of the process of inclusion/exclusion of patients data included in the FEPSY database.

excluded those for whom the birth year was missing or equal to the year of the first visit to mental health services, as it indicates that there was an error during the registration of the patient. The resulting dataset included socio-demographic (e.g., nationality, birth area, residence area, . . .) and clinical (e.g., compulsory and voluntary hospitalizations, total days of treatment, received diagnoses other than psychosis, . . .) data of 39,241 individuals. The dataset generation process is summarized in Fig. 1.

The features included in the dataset and used for the classification experiments are listed in Table 1. In order to predict a diagnosis of psychotic spectrum disorders, users were labeled with “Yes” if they had received a diagnosis of psychosis at least once during the observation period, and with “No” otherwise. As previously reported, and based on the international incidence of psychotic disorders [27], we expect the dataset to be imbalanced regarding diagnosis. As a matter of fact, only the 7.06% of the users had psychosis and the IR is 13.166. The resulting labeled dataset is therefore considered imbalanced.

To perform classification experiments, the dataset was divided in two parts: the training set, including 70% of the original dataset, and the test set, including the remaining 30%. Class distributions were preserved as observed in the original dataset.

3.2. Balancing techniques

In order to solve the class imbalance problem, different training sets with different percentages of class imbalance, namely 50%–50%, 60%–40%, and 70%–30% were created. `Imbalanced-learn` [28] is an open-source library for Python programming language, that provides various sampling techniques to use when working with imbalanced datasets. `Imbalanced-learn`’s *RandomOverSampler* (ROS), *SMOTE-NC*, and *RandomUnderSampler* (RUS) methods were applied to the original training set to reduce the IR.

3.3. Machine learning methods

All the classification experiments were conducted with the Waikato Environment for Knowledge Analysis (Weka) [29,30] data mining tool, an open-source software developed by the University of Waikato, in New Zealand, which provides several algorithms and tools for data preprocessing and visualization, classification, clustering, and feature selection.

For classification experiments, different Weka’s classifiers were chosen:

- *Logistic*, a multinomial logistic regression model with a ridge estimator [31];
- *J48*, i.e. Weka’s implementation of the C4.5 algorithm [32] to build a decision tree;
- *RandomForest* [33], an ensemble learning method that builds multiple decision trees and assigns the label returned by the majority of the trees;
- *AdaBoostM1*, a boosting algorithm [34];
- *PART* [35], that builds a decision list, i.e. an ordered set of rules expressed in the form of IF-THEN rules, that together form a classifier [36,37];
- *Vote*, which builds a voting classifier by combining the classifiers listed above [38,39].

The classifiers were first trained on the different training sets and then tested on the imbalanced test set.

At the end of the training, an example falls in exactly one of four cases:

- True Positive (TP): a positive example classified as positive
- True Negative (TN): a negative example classified as negative
- False Positive (FP): a negative example classified as positive
- False Negative (FN): a positive example classified as negative

The result of the classification can be summarized in a confusion matrix, which gathers the numbers of TP, TN, FP, and FN, as shown in Fig. 2.

We evaluated the classifiers in terms of:

- *accuracy*, the proportion of correct predictions:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Table 1
Demographic and clinical characteristics of the individuals included in this study.

Variable	Total (%)	Non-Psychosis (%)	Psychosis (%)
	(n=39,241)	(n=36,471)	(n=2,770)
Sex			
Female	24140 (61.52)	22673 (93.92)	1467 (6.08)
Male	15101 (38.48)	13798 (91.37)	1303 (8.63)
Born in/outside Italy			
Italy	36189 (92.22)	33752 (92.54)	2437 (87.98)
Outside Italy	2917 (7.43)	2607 (7.15)	310 (11.19)
Unknown	135 (0.34)	112 (0.31)	23 (0.83)
Birth Area			
Other province	10480 (26.71)	9688 (26.56)	792 (28.59)
Ferrara	10391 (26.48)	9638 (26.43)	753 (27.18)
Codigoro	4804 (12.24)	4478 (12.28)	326 (11.77)
Portomaggiore	3877 (9.88)	3671 (10.07)	206 (7.44)
Copparo	3509 (8.94)	3341 (9.16)	168 (6.06)
Cento	3128 (7.97)	2936 (8.05)	192 (6.93)
Outside Italy	2917 (7.43)	2607 (7.15)	310 (11.19)
Unknown	135 (0.34)	112 (0.31)	23 (0.83)
Residence Area			
Ferrara	14410 (36.72)	13446 (36.87)	964 (34.80)
Codigoro	5901 (15.04)	5465 (14.98)	436 (15.74)
Other province	4622 (11.78)	4127 (11.32)	495 (17.87)
Cento	4511 (11.50)	4278 (11.73)	233 (8.41)
Portomaggiore	4507 (11.49)	4270 (11.71)	237 (8.56)
Copparo	3948 (10.06)	3755 (10.30)	193 (6.97)
Unknown	984 (2.51)	838 (2.30)	146 (5.27)
Outside Italy	358 (0.91)	292 (0.80)	66 (2.38)
Age at first visit			
<18	266 (0.68)	248 (0.68)	18 (0.65)
18–24	3144 (8.01)	2914 (7.99)	230 (8.30)
25–34	5796 (14.77)	5256 (14.41)	540 (19.49)
35–44	7140 (18.20)	6576 (18.03)	564 (20.36)
45–54	6731 (17.15)	6208 (17.02)	523 (18.88)
55–64	5812 (14.81)	5370 (14.72)	442 (15.96)
65–74	5275 (13.44)	4969 (13.62)	306 (11.05)
75+	5077 (12.94)	4930 (13.52)	147 (5.31)
Age at discharge			
<18	146 (0.37)	143 (0.39)	3 (0.11)
18–24	1688 (4.30)	1606 (4.40)	82 (2.96)
25–34	4291 (10.93)	4005 (10.98)	286 (10.32)
35–44	6737 (17.17)	6278 (17.21)	459 (16.57)
45–54	6944 (17.70)	6450 (17.69)	494 (17.83)
55–64	5962 (15.19)	5522 (15.14)	440 (15.88)
65–74	5620 (14.32)	5135 (14.08)	485 (17.51)
75+	7853 (20.01)	7332 (20.10)	521 (18.81)
Total no. of diagnoses			
(mean ± sd [min;max])	2.03 ± 2.67 [0;121]	1.87 ± 2.35 [0;121]	4.13 ± 4.83 [1;60]
No. of distinct diagnoses			
(mean ± sd [min;max])	1.14 ± 0.76 [0;7]	1.10 ± 0.71 [0;7]	1.77 ± 1.01 [1;7]
Anxiety disorders			
no	26770 (68.22)	24371 (66.82)	2399 (86.61)
yes	12471 (31.78)	12100 (33.18)	371 (13.39)
Depression			
no	35120 (89.50)	32658 (89.55)	2462 (88.88)
yes	4121 (10.50)	3813 (10.45)	308 (11.12)
Drug and substance use/abuse			
no	37529 (95.64)	34904 (95.70)	2625 (94.77)
yes	1712 (4.36)	1567 (4.30)	145 (5.23)
Eating disorders			
no	38953 (99.27)	36190 (99.23)	2763 (99.75)
yes	288 (0.73)	281 (0.77)	7 (0.25)
Intellectual Disability			
no	38124 (97.15)	35478 (97.28)	2646 (95.52)
yes	1117 (2.85)	993 (2.72)	124 (4.48)
Mania and Bipolar Disorders			
no	37737 (96.17)	35180 (96.46)	2557 (92.31)
yes	1504 (3.83)	1291 (3.54)	213 (7.69)
Organic Psychosis			
no	35120 (89.50)	32658 (89.55)	2462 (88.88)
yes	4121 (10.50)	3813 (10.45)	308 (11.12)
Personality Disorders			
no	35992 (91.72)	33549 (91.99)	2443 (88.19)
yes	3249 (8.28)	2922 (8.01)	327 (11.81)
Infantile autism			
no	39177 (99.84)	36439 (99.91)	2738 (98.84)
yes	64 (0.16)	32 (0.09)	32 (1.16)

(*continued on next page*)

Table 1 (*continued*).

Variable	Total (%) (n=39,241)	Non-Psychosis (%) (n=36,471)	Psychosis (%) (n=2,770)
Other mental disorders			
no	34892 (88.92)	32213 (88.32)	2679 (96.71)
yes	4349 (11.08)	4258 (11.68)	91 (3.29)
No. of visits			
1	27857 (70.99)	26524 (72.73)	1333 (48.12)
2+	10253 (26.13)	9099 (24.95)	1154 (41.66)
5+	1131 (2.88)	848 (2.33)	283 (10.22)
No. of hospitalizations			
0	35798 (91.23)	34249 (93.91)	1549 (55.92)
1	2138 (5.45)	1491 (4.09)	647 (23.36)
2+	1305 (3.33)	731 (2.00)	574 (20.72)
Avg. duration of hosp.			
Never hospitalized	35798 (91.23)	34249 (93.91)	1549 (55.92)
1-7 days	1423 (3.63)	1046 (2.87)	377 (13.61)
> 7 days	2020 (5.15)	1176 (3.22)	844 (30.47)
Had compulsory psychiatric hospitalization			
no	38780 (98.83)	36267 (99.44)	2513 (90.72)
yes	461 (1.17)	204 (0.56)	257 (9.28)
Total duration of treatment (in days)			
1–15	10324 (26.31)	10031 (27.50)	293 (10.58)
>15	1638 (4.17)	1567 (4.30)	71 (2.56)
>30	2527 (6.44)	2462 (6.75)	65 (2.35)
>60	1296 (3.30)	1264 (3.47)	32 (1.16)
>90	2600 (6.63)	2504 (6.87)	96 (3.47)
>180	2665 (6.79)	2562 (7.02)	103 (3.72)
>365	4441 (11.32)	4120 (11.30)	321 (11.59)
>1095	3040 (7.75)	2736 (7.50)	304 (10.97)
>1825	4685 (11.94)	4146 (11.37)	539 (19.46)
>3650	6025 (15.35)	5079 (13.93)	946 (34.15)

- *precision*, the proportion of positive examples found among all those classified as positive:

$$PPV = \frac{TP}{TP + FP}$$

- *recall* (*sensitivity*, *true positive rate*), the proportion of positive examples found among all the positive ones:

$$TPR = \frac{TP}{TP + FN}$$

- specificity (*true negative rate*) is the proportion of negative examples found among all the negative ones:

$$TNR = \frac{TN}{TN + FP}$$

- *F1-score*, the harmonic mean of precision and recall:

$$F1\text{-}score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

- *balanced accuracy*, the arithmetic mean of sensitivity and specificity (the proportion of negative examples found among all the negative ones):

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2}$$

In this study, we focused especially on the recall. Since a high value of the recall reflects a low number of false negatives, we were more interested in correctly classifying patients with psychosis, that is, decreasing the number of false negatives. We also took into consideration the F1-score, as it reflects both the ability of the model to find the majority of all positive examples (recall) and its ability to identify only positive examples (precision). Ideally, a model should have a high F1-score, meaning high precision and recall. However, in real-world applications the higher the precision, the lower the recall (or vice versa), and a model is tuned based on its application. Having an imbalanced dataset, we were also interested in balanced accuracy, which is a more reliable metric than accuracy in this scenario. In fact, taking into account both the true positive rate and the true negative rate gives a better idea on how well the model is able to predict both classes.

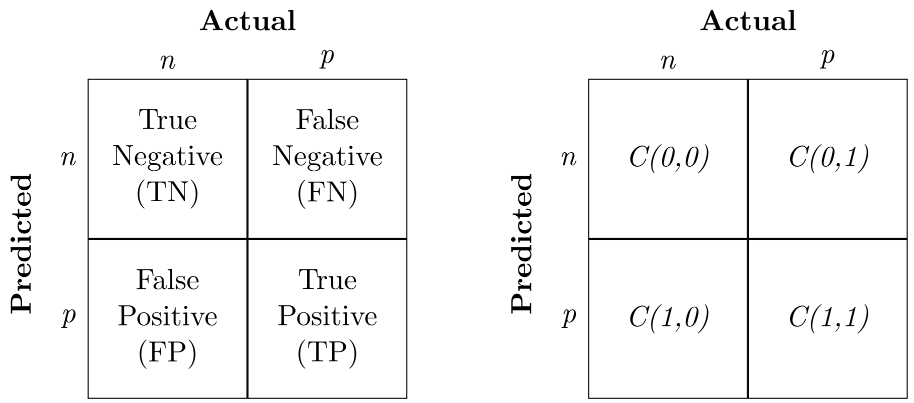
3.4. Cost-sensitive learning

Most ML algorithms assume not only that class distribution is the same, but also that misclassification errors (i.e., false negative and false positive costs) are equally important [18].

In real-world settings, and especially in healthcare, this is not true: the cost of misclassifying an ill patient is greatly higher than the cost of misclassifying a healthy patient [19]. Cost-sensitive learning techniques have been widely employed in the medical field for diagnosis [6–8,20–23], but other examples can be found in other fields as well, such as fraud detection [3,40,41], manufacturing for predicting product failures [42], intrusion detection [4,5,43], and anomaly detection [44]. In each of these fields, the class of interest is always the underrepresented one.

Cost-sensitive learning refers to those algorithm-level techniques in which the goal of the training is to minimize the cost of a model on a training set. They are especially used to deal with class imbalance. Models are trained while taking into account a cost matrix. In the notation introduced by [12], a cost matrix C has the same structure as a confusion matrix, and each cell $C(i,j)$ represents the cost of predicting class i when the actual class is j. For example, if we assign 0 to the negative label and 1 to the positive label, then $C(0,0)$ is the cost of predicting a negative label when the true one is negative. The structure of a generic cost matrix is depicted in Fig. 2.

In binary classification, the default cost matrix has a cost of 0 for correct predictions ($i = j$) and a cost of 1 for incorrect ones ($i \neq j$). By modifying cost values it is possible to assign a greater weight to one or another type of error. Since defining the cost matrix is a challenging task (starting from the definition of cost) and costs are hardly ever known in real applications, a good starting point consists in assigning the number of instances of the opposite class [19], or equivalently the inverse class distribution ratio. We confronted the performances on the original dataset with two Cost-sensitive learning algorithms included



(a) Structure of a confusion matrix, where p stands for positive and n stands for negative.

(b) Structure of a cost matrix, where $C(i,j)$ is the cost of predicting class i when the actual class is j.

Fig. 2. Structures of a confusion matrix (a) and of a cost matrix (b).

in Weka: MetaCost [45] and CosSensitiveClassfier [30]. As stated in Weka, MetaCost builds an ensemble classifier using bagging and uses it to label each training instance with the prediction that minimizes the expected cost, based on the probability estimates obtained from bagging. If an interpretable base learner is chosen, then the output will be interpretable as well. On the other hand, CosSensitiveClassfier makes its base classifier cost-sensitive, using two methods: predicting the class with minimum expected misclassification cost (CSC) (done by setting the attribute *minimizeExpectedCost* = *true*); reweighting training instances according to the total cost assigned to each class (CSC-W) (*minimizeExpectedCost* = *false*).

In the first case, we actually perform cost-sensitive classification, which adjusts the output of the classifier to optimize the given cost matrix; in this case, costs are ignored at training time and used only at prediction time. Conversely, with reweighting the cost matrix is taken into account during training and ignored at prediction time. We applied Bagging, an ensemble method that improves the performances of the base classifier [45], with 10 iterations and used J48 as base classifier. We first tested the above-mentioned cost-sensitive learning techniques with a cost ratio of 1:13.166, where 13.166 was the IR of the dataset. Then, in order to make the ratio independent of the IR, we tried also other ratios: 1:5, 1:10, and 1:15.

3.5. Feature selection

With the term feature selection (or attribute selection) we refer to the automatic process of selecting a subset of relevant features (attributes, variables, predictors) used in the construction of the predictive model. It is a fundamental process of ML algorithms as it helps (a) to remove unnecessary, irrelevant and/or distracting variables that will not help increase the model accuracy (they might in fact deteriorate it); (b) to find more effective predictors; and (c) to reduce the model complexity, making it more comprehensible by humans [46]. In order to investigate the relevance of each feature of the dataset used in this study, we conducted a feature selection analysis. Four different Weka's feature evaluators were employed to assign a score to each feature, and the Ranker method to create a ranking based on that score. We then averaged these scores to find the most relevant features. Below is the list of the selected evaluators:

- *CorrelationAttributeEval*, which evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class;
- *GainRatioAttributeEval*, which scores attributes by measuring their gain ratio with respect to the class;
- *SymmetricalUncertAttributeEval*, which evaluates the worth of an attribute by measuring the symmetrical uncertainty with respect to the class.

4. Experimental results

Table 2 shows the average performances of the selected classifiers on the original and on the resampled training sets. Although we focused on balanced accuracy, recall, and F1-score during the experiments, Table 2 also shows accuracy to make clear the metric's lack of usefulness in the case of imbalanced datasets, precision to show the trade-off with recall (intercepted by F1-score), and sensitivity, used to calculate the balanced accuracy. On the original training set, we obtained relatively high balanced accuracy and precision. All the classifiers seemed to be able to distinguish quite well patients without psychosis. By applying the different balancing techniques, all the classifiers were in fact able to identify a higher number of patients with psychosis (i.e., true positives), leading to an increase in the recall at the price of a lower precision (i.e., a higher number of false positives). In some cases, the F1-score improved as well, although remaining around 0.5.

Table 3 reports in detail the balanced accuracy, recall, and F1-score obtained on the test set by each classifier trained on the different training sets. The Logistic regression trained on 50-50 ROS achieved both the highest balanced accuracy and recall, while the voting classifier trained on the 70-30 SMOTE-NC achieved the highest F1-score.

In Table 4 we reported the training times (in seconds) taken to build each model on each training set. Being the simplest algorithm, J48 was always the fastest; on the other hand, Vote required a significantly higher amount of time to train, having to train different models inside to take their votes. The RUS setting provided the lowest times, building the smaller datasets. On the contrary, ROS and SMOTENC were reasonably the most time-consuming to train on. Despite all, training times were always under 20 s, and thus even the slowest technique could be applied in scenarios that do not require real-time decisions.

Table 5 shows the results of the three algorithms we used for cost-sensitive learning with different cost ratios, obtained on the test set. Using the IR proved to be a good choice, leading each model to high values of both recall and balanced accuracy with all algorithms. However, the setting that obtained the highest values was the CSC with the 1:15 cost ratio. Nonetheless, ratios with lower costs for false positives errors achieved comparable results. Training times were relatively higher than those of the single ML algorithms, due to the fact that we used 100 iterations for the bagging.

Table 6 shows the average scores of the features obtained on the original and resampled training sets using Weka's attribute evaluators. Hospitalizations (compulsory and voluntary) and number of diagnoses (distinct and total) appear to be the most relevant ones. On the other hand, socio-demographic features such as birth area and residence area, age at first visit and at discharge do not seem to affect much the classification.

Table 2
Average classification results obtained on the test set, using Weka's classifiers trained on the original and on the resampled training sets.

	Training	Average metric					
	dataset	Balanced accuracy	Accuracy	Precision	Recall	F1-score	Specificity
	Original	0.686	0.946	0.719	0.383	0.496	0.912
50–50	RUS	0.856	0.846	0.302	0.868	0.447	0.929
	ROS	0.819	0.906	0.419	0.718	0.520	0.899
	SMOTE-NC	0.808	0.904	0.407	0.696	0.508	0.895
60–40	RUS	0.847	0.880	0.355	0.809	0.492	0.930
	ROS	0.809	0.917	0.447	0.684	0.536	0.897
	SMOTE-NC	0.799	0.920	0.456	0.658	0.536	0.899
70–30	RUS	0.826	0.911	0.432	0.725	0.539	0.930
	ROS	0.792	0.929	0.498	0.632	0.554	0.900
	SMOTE-NC	0.785	0.929	0.501	0.618	0.551	0.904

Table 3
Balanced Accuracy, Recall, and F1-score obtained on the test set, with each Weka's classifier trained on the original and on the sampled training sets.

	Training	Balanced accuracy					
	dataset	*AdaBoostM1*	*J48*	*Logistic*	*PART*	*RandomForest*	*Vote*
	Original	0.626	0.683	0.724	0.732	0.658	0.693
50–50	RUS	0.849	0.853	0.890	0.853	0.818	0.876
	ROS	0.853	0.795	**0.892**	0.797	0.743	0.833
	SMOTE-NC	0.828	0.789	0.863	0.789	0.759	0.819
60–40	RUS	0.828	0.851	0.884	0.839	0.817	0.867
	ROS	0.832	0.797	0.884	0.781	0.742	0.818
	SMOTE-NC	0.809	0.783	0.857	0.785	0.753	0.808
70–30	RUS	0.786	0.830	0.862	0.827	0.804	0.845
	ROS	0.784	0.795	0.862	0.778	0.730	0.800
	SMOTE-NC	0.786	0.782	0.836	0.764	0.741	0.803

	Training	Recall					
	dataset	*AdaBoostM1*	*J48*	*Logistic*	*PART*	*RandomForest*	*Vote*
	Original	0.258	0.375	0.460	0.485	0.327	0.392
50-50	RUS	0.821	0.887	0.909	0.864	0.834	0.897
	ROS	0.819	0.664	**0.913**	0.655	0.532	0.724
	SMOTE-NC	0.787	0.645	0.830	0.644	0.575	0.697
60-40	RUS	0.734	0.836	0.863	0.807	0.779	0.838
	ROS	0.747	0.664	0.859	0.620	0.528	0.687
	SMOTE-NC	0.692	0.627	0.789	0.626	0.554	0.663
70-30	RUS	0.621	0.742	0.787	0.744	0.704	0.755
	ROS	0.616	0.649	0.786	0.606	0.497	0.638
	SMOTE-NC	0.625	0.615	0.729	0.572	0.525	0.640

	Training	F1-score					
	dataset	*AdaBoostM1*	*J48*	*Logistic*	*PART*	*RandomForest*	*Vote*
	Original	0.389	0.497	0.567	0.553	0.442	0.530
50–50	RUS	0.476	0.416	0.503	0.437	0.376	0.471
	ROS	0.495	0.505	0.506	0.535	0.500	0.578
	SMOTE-NC	0.450	0.511	0.518	0.514	0.492	0.562
60–40	RUS	0.529	0.464	0.553	0.459	0.422	0.523
	ROS	0.527	0.515	0.564	0.520	0.503	0.585
	SMOTE-NC	0.517	0.518	0.565	0.533	0.505	0.580
70–30	RUS	0.547	0.523	0.603	0.507	0.475	0.579
	ROS	0.547	0.538	0.604	0.532	0.503	0.599
	SMOTE-NC	0.538	0.539	0.589	0.530	0.504	**0.609**

5. Discussion

This study explored the applicability of a ML model to identify patients with a diagnosis of psychotic spectrum disorders using real-world EHR data. The extracted dataset was heavily imbalanced, with only 7.06% of patients with at least one diagnosis of psychosis (IR = 13.166). The different balancing techniques we adopted showed that they can effectively help to find the best setting to accomplish the classification task. Balancing the dataset led to a substantial improvement of the recall: in the best case, we could achieve a balanced accuracy of around 80% and a recall of around 90%. Even though the improvement of the F1-score was negligible and the precision decreased, the results we obtained were far better compared to those obtained on the original dataset, since we were more interested in correctly classifying patients

Table 4
Training times to build each model on each training set, in seconds.

	Training dataset	Classifier *AdaBoostM1*	*J48*	*Logistic*	*PART*	*RandomForest*	*Vote*
	Original	1.91	0.19	2.55	0.67	1.85	6.64
50–50	RUS	0.29	0.04	0.35	0.11	0.44	1.07
	ROS	3.87	0.36	4.89	4.88	3.19	16.89
	SMOTE-NC	4.11	0.44	5.10	3.94	3.61	16.84
60–40	RUS	0.37	0.04	0.44	0.11	0.50	1.48
	ROS	3.21	0.26	4.00	3.51	2.69	13.88
	SMOTE-NC	3.25	0.31	4.19	2.70	2.98	13.27
70–30	RUS	0.47	0.09	0.56	0.16	0.60	1.77
	ROS	2.69	0.24	3.40	2.75	2.48	11.56
	SMOTE-NC	2.70	0.28	3.55	2.07	2.70	11.12

Table 5
Cost-sensitive learning results obtained on the test set. For each metric of interest (balanced accuracy, recall, F1-score, and training time), the best result is written in bold.

	Cost ratio	Balanced accuracy	Accuracy	Precision	Recall	F1-score	TNR	Training time
CSC	1:5	0.811	0.932	0.515	0.670	0.582	0.952	**11.54**
	1:10	0.837	0.904	0.404	0.759	0.528	0.915	11.97
	1:13.166	0.840	0.876	0.339	0.798	0.476	0.882	11.81
	1:15	**0.842**	0.869	0.328	**0.811**	0.467	0.874	11.67
CSC-W	1:5	0.814	0.932	0.516	0.676	**0.585**	0.952	15.87
	1:10	0.828	0.914	0.434	0.728	0.544	0.928	15.43
	1:13.166	0.835	0.905	0.407	0.752	0.529	0.917	14.75
	1:15	0.835	0.900	0.392	0.759	0.517	0.911	15.42
MetaCost	1:5	0.779	0.935	0.537	0.597	0.566	0.961	12.14
	1:10	0.832	0.911	0.426	0.739	0.541	0.925	12.82
	1:13.166	0.834	0.882	0.350	0.779	0.482	0.890	12.51
	1:15	0.833	0.873	0.331	0.786	0.466	0.880	11.93

Table 6
Average scores of the features resulting from the evaluation on the original and resampled training sets with Weka's attribute evaluators.

Feature	Average score
Avg. duration of hospitalization	0.21960
No. of hospitalization	0.21613
Had compulsory psychiatric hospitalization	0.11560
No. of distinct diagnoses	0.10384
Anxiety disorder	0.10299
Total no. of diagnoses	0.09776
No. of visits	0.09343
Other mental disorders	0.09314
Depression	0.07451
Total duration of treatment	0.06669
Infantile autism	0.04008
Sex	0.03524
Mania and Bipolar Disorders	0.02779
Eating disorders	0.02698
Born in/outside Italy	0.02400
Intellectual Disability	0.01781
Residence Area	0.01760
Age at first visit	0.01703
Personality Disorders	0.01461
Birth Area	0.01451
Drug and substance use/abuse	0.01174
Age at discharge	0.00789
Organic Psychosis	0.00764

with psychosis, that is, decreasing the number of false negatives. In general, having a low number of false negatives is especially important and preferable in healthcare: it is better to label a healthy patient as "ill" (eventually a more thorough diagnosis will prove it wrong) rather than to label an ill patient as "healthy" [47]. Therefore, the goal was to decrease the number of false negatives (and consequently increase the number of true positives). Among all the balancing techniques, 50-50 ROS appeared to be the best balancing approach in this case. Overall, none of the classifiers significantly outperformed the others.

The feature analysis suggested that clinical features play a more relevant role in the classification rather than socio-demographic ones.

The ML models we tried, the evaluation metrics we chose, and the results we obtained were comparable to those listed in Chung and Teo's [48] review of ML approaches for mental health prediction.

Aside from good performances, recently there has also been an increasing interest in interpretable ML models, that is, models whose output is easily understood by humans [49]. Interpretability is especially important in those fields where the decision of a ML model needs to be trusted, such as healthcare [47,50]. In this regard, three of the models we chose, namely logistic regression, decision list and decision tree, are known for being easily understood also by non-experts [49]. The logistic regression is also the model that overall performed better.

After selecting one or an ensemble of ML models and the most suitable balancing technique, if the ML model is sufficiently reliable in its performances, an approach built with this methodology could be employed to assist clinicians and researchers in detecting and diagnosing mental health disorders and improve treatment [48,51–53]. In the absence of clear differences in the performances, if interpretability is important for the application, the decision tree (J48) could be employed, being the most interpretable and fastest.

As mentioned in the introduction, ML often deals with synthetic experiments considering ideal dataset balancing conditions, as each class has about the same number of examples in a binary or multiple classification context. When datasets originate from industrial or medical applications, the scenario is often very different and more challenging. First, very frequently these datasets were not created with the intent of using them in learning from examples contexts, therefore they need to be cleaned up and preprocessed before they are ready for real use in a ML context. Second, since these data were not collected for classification purposes, they must be labeled according to the question they are intended to answer. The medical context, along with anomaly detection context, is by far one of the major contexts in which datasets are usually particularly imbalanced. The application of the above-mentioned balancing techniques can improve the performance in the training phase; on the other hand, modifying the nature of the dataset can worsen the results in the testing phase. Cost-sensitive learning does not alter the dataset by adding or removing instances, but instead it assigns costs to the different types of error, trying to penalize more errors on the class of interest. In fact, it is already successfully applied in such scenarios.

6. Conclusion

In this work, we presented a classification task applied to a real-world mental health imbalanced dataset. We compared the results of both data-level and algorithm-level techniques, obtained using several classifiers of the Weka data mining tool. The presented findings suggest that the application of both ML and balancing techniques to an imbalanced dataset may be useful for the prediction of a specific diagnosis. However, before the described approach could be employed in routine clinical practice, more studies are needed to improve the techniques for treating imbalanced datasets to increase their reliability. The obtained results can be considered a good starting point for future works, which will aim to improve also the precision. Other sampling techniques, such as generative models, can be investigated as well.

Funding

This work was supported by the “Fondo per l’Incentivazione alla Ricerca (FIR)” year titled “Intelligenza Artificiale per Predizione Diagnostica, di Carico Assistenziale, e di Esito: 40 Anni di accessi Presso i Centri di Salute Mentale della Provincia di Ferrara (AI4MentalHealth)”, granted by the University of Ferrara in 2021 to Maria Ferrara; and partly supported by the “National Group of Computing Science (GNCS-INdAM)”. The publication was created with the co-financing of the European Union-FSE-REACT-EU, PON Research and Innovation 2014–2020 DM1062/2021.

This publication was produced while attending the PhD programme in Engineering Science at the University of Ferrara, Cycle XXXVIII, with the support of a scholarship financed by the Ministerial Decree no. 351 of 9th April 2022, based on the NRRP - funded by the European Union - NextGenerationEU - Mission 4 “Education and Research”, Component 1 “Enhancement of the offer of educational services: from nurseries to universities” - Investment 4.1 “Extension of the number of research doctorates and innovative doctorates for public administration and cultural heritage”.

CRediT authorship contribution statement

Elisabetta Gentili: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Giorgia Franchini:** Conceptualization, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. **Riccardo Zese:** Conceptualization, Formal analysis, Methodology, Writing – review & editing. **Marco Alberti:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Maria Ferrara:** Conceptualization, Funding acquisition, Writing – review & editing. **Ilaria Domenicano:** Data curation, Writing – review & editing. **Luigi Grassi:** Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

All authors disclose they did not have any financial or personal relationships with other people or organizations that could inappropriately influence (bias) their work.

References

[1] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284, http://dx.doi.org/10.1109/TKDE.2008.239.

[2] B. Krawczyk, Learning from imbalanced data: Open challenges and future directions, Prog. Artif. Intell. 5 (4) (2016) 221–232, http://dx.doi.org/10.1007/s13748-016-0094-0.

[3] S. Makki, Z. Assaghir, Y. Taher, R. Haque, M.-S. Hacid, H. Zeineddine, An experimental study with imbalanced classification approaches for credit card fraud detection, IEEE Access 7 (2019) 93010–93022.

[4] G. Karatas, O. Demir, O.K. Sahingoz, Increasing the performance of machine learning-based IDSs on an imbalanced and up-to-date dataset, IEEE Access 8 (2020) 32150–32162.

[5] L. Liu, P. Wang, J. Lin, L. Liu, Intrusion detection of imbalanced network traffic based on machine learning and deep learning, IEEE Access 9 (2020) 7550–7563.

[6] M. Khalilia, S. Chakraborty, M. Popescu, Predicting disease risks from highly imbalanced data using random forest, BMC Med. Inform. Decis. Making 11 (1) (2011) 1–13.

[7] A. Majid, S. Ali, M. Iqbal, N. Kausar, Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines, Comput. Methods Programs Biomed. 113 (3) (2014) 792–808.

[8] W. Gerych, E. Agu, E. Rundensteiner, Classifying depression in imbalanced datasets using an autoencoder-based anomaly detection approach, in: 2019 IEEE 13th International Conference on Semantic Computing, ICSC, 2019, pp. 124–127, http://dx.doi.org/10.1109/ICOSC.2019.8665535.

[9] Y. Sun, A.K. Wong, M.S. Kamel, Classification of imbalanced data: A review, Int. J. Pattern Recogn. Artif. Intell. 23 (04) (2009) 687–719.

[10] P. Branco, L. Torgo, R.P. Ribeiro, A survey of predictive modeling on imbalanced domains, ACM Comput. Surv. (CSUR) 49 (2) (2016) 1–50.

[11] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.

[12] C. Elkan, The foundations of cost-sensitive learning, in: International Joint Conference on Artificial Intelligence. Vol. 17, Lawrence Erlbaum Associates Ltd, 2001, pp. 973–978.

[13] M.M. Rahman, D.N. Davis, Addressing the class imbalance problem in medical datasets, Int. J. Mach. Learn. Comput. (2013) 224–228.

[14] S. Belarouci, M.A. Chikh, Medical imbalanced data classification, Adv. Sci. Technol. Eng. Syst. J. 2 (3) (2017) 116–124.

[15] D.-C. Li, C.-W. Liu, S.C. Hu, A learning method for the class imbalance problem with medical data sets, Comput. Biol. Med. 40 (5) (2010) 509–518, http://dx.doi.org/10.1016/j.compbiomed.2010.03.005.

[16] M. Khushi, K. Shaukat, T.M. Alam, I.A. Hameed, S. Uddin, S. Luo, X. Yang, M.C. Reyes, A comparative performance analysis of data resampling methods on imbalanced medical data, IEEE Access 9 (2021) 109960–109975, http://dx.doi.org/10.1109/ACCESS.2021.3102399.

[17] M. Zeng, B. Zou, F. Wei, X. Liu, L. Wang, Effective prediction of three common diseases by combining SMOTE with tomek links technique for imbalanced medical data, in: 2016 IEEE International Conference of Online Analysis and Computing Science, ICOACS, 2016, pp. 225–228, http://dx.doi.org/10.1109/ICOACS.2016.7563084.

[18] N. Thai-Nghe, Z. Gantner, L. Schmidt-Thieme, Cost-sensitive learning methods for imbalanced data, in: The 2010 International Joint Conference on Neural Networks, IJCNN, 2010, pp. 1–8, http://dx.doi.org/10.1109/IJCNN.2010.5596486.

[19] V.S. Sheng, C.X. Ling, Thresholding for making classifiers cost-sensitive, in: Aaai. Vol. 6, 2006, pp. 476–481.

[20] I.D. Mienye, Y. Sun, Performance analysis of cost-sensitive learning methods with application to imbalanced medical data, Inform. Med. Unlocked 25 (2021) 100690.

[21] M. Zhu, J. Xia, X. Jin, M. Yan, G. Cai, J. Yan, G. Ning, Class weights random forest algorithm for processing class imbalanced medical data, IEEE Access 6 (2018) 4641–4652, http://dx.doi.org/10.1109/ACCESS.2018.2789428.

[22] M. Phankokkruad, Cost-sensitive extreme gradient boosting for imbalanced classification of breast cancer diagnosis, in: 2020 10th IEEE International Conference on Control System, Computing and Engineering, ICCSCE, 2020, pp. 46–51, http://dx.doi.org/10.1109/ICCSCE50387.2020.9204948.

[23] S.I. Ali, H.S.M. Bilal, M. Hussain, J. Hussain, F.A. Satti, M. Hussain, G.H. Park, T. Chung, S. Lee, Ensemble feature ranking for cost-based non-overlapping groups: A case study of chronic kidney disease diagnosis in developing countries, IEEE Access 8 (2020) 215623–215648, http://dx.doi.org/10.1109/ACCESS.2020.3040650.

[24] M. Ferrara, E. Gentili, M. Belvederi Murri, R. Zese, M. Alberti, G. Franchini, I. Domenicano, F. Folesani, C. Sorio, L. Benini, P. Carozza, J. Little, L. Grassi, Establishment of a public mental health records registry in the Ferrara province: Adaptation of a 30-year-long clinical database for research purposes, 2023, Manuscript under submission.

[25] A. Fioritti, L. Lo Russo, V. Melega, Reform said or done? The case of Emilia-Romagna within the Italian psychiatric context, Am. J. Psychiatry 154 (1) (1997) 94–98.

[26] World Health Organization, International Classification of Diseases : [9th] Ninth Revision, Basic Tabulation List with Alphabetic Index, World Health Organization, 1978, p. 331.

[27] H.E. Jongsma, C. Turner, J.B. Kirkbride, P.B. Jones, International incidence of psychotic disorders, 2002–17: A systematic review and meta-analysis, Lancet Public Health 4 (5) (2019) e229–e244.

[28] G. Lemaître, F. Nogueira, C.K. Aridas, Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning, J. Mach. Learn. Res. 18 (17) (2017) 1–5, URL http://jmlr.org/papers/v18/16-365.html.

[29] E. Frank, M.A. Hall, I.H. Witten, Online appendix for data mining: Practical machine learning tools and techniques, Data Mining: Practical Machine Learning Tools and Techniques, Fourth Ed., Morgan Kaufmann, 2016.

[30] I.H. Witten, E. Frank, M.A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, third ed., Morgan Kaufmann, 2011.

[31] S.L. Cessie, J.V. Houwelingen, Ridge estimators in logistic regression, J. R. Stat. Soc. Ser. C. Appl. Stat. 41 (1) (1992) 191–201.

[32] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, 1993.

[33] T.K. Ho, Random decision forests, in: Proceedings of 3rd International Conference on Document Analysis and Recognition. Vol. 1, 1995, pp. 278–282, http://dx.doi.org/10.1109/ICDAR.1995.598994.

[34] Y. Freund, R.E. Schapire, Experiments with a new boosting algorithm, in: Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, ICML '96, Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, 1996, pp. 148–156.

[35] E. Frank, I.H. Witten, Generating accurate rule sets without global optimization, in: J. Shavlik (Ed.), Fifteenth International Conference on Machine Learning, Morgan Kaufmann, 1998, pp. 144–151.

[36] R.L. Rivest, Learning decision lists, Mach. Learn. 2 (1987) 229–246.

[37] J. Fürnkranz, T. Kliegr, A brief overview of rule learning, in: Rule Technologies: Foundations, Tools, and Applications: 9th International Symposium, RuleML 2015, Berlin, Germany, August 2-5, 2015, Proceedings 9, Springer, 2015, pp. 54–69.

[38] L.I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, John Wiley & Sons, 2014.

[39] J. Kittler, M. Hatef, R.P. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 20 (3) (1998) 226–239.

[40] P.K. Chan, S.J. Stolfo, Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection, in: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, KDD '98, AAAI Press, 1998, pp. 164–168.

[41] Y. Sahin, S. Bulkan, E. Duman, A cost-sensitive decision tree approach for fraud detection, Expert Syst. Appl. 40 (15) (2013) 5916–5923.

[42] F.D. Frumosu, A.R. Khan, H. Schiøler, M. Kulahci, M. Zaki, P. Westermann-Rasmussen, Cost-sensitive learning classification strategy for predicting product failures, Expert Syst. Appl. 161 (2020) 113653.

[43] W. Lee, W. Fan, M. Miller, S.J. Stolfo, E. Zadok, Toward cost-sensitive modeling for intrusion detection and response, J. Comput. Secur. 10 (1–2) (2002) 5–22.

[44] H.-y. Lu, F.-y. Chen, M. Xu, C.-j. Wang, J.-y. Xie, Never ignore the significance of different anomalies: A cost-sensitive algorithm based on loss function for anomaly detection, in: 2015 IEEE 27th International Conference on Tools with Artificial Intelligence, ICTAI, 2015, pp. 1099–1105, http://dx.doi.org/10.1109/ICTAI.2015.156.

[45] P. Domingos, Metacost: A general method for making classifiers cost-sensitive, in: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999, pp. 155–164.

[46] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (Mar) (2003) 1157–1182.

[47] M. Mozaffari-Kermani, S. Sur-Kolay, A. Raghunathan, N.K. Jha, Systematic poisoning attacks on and defenses for machine learning in healthcare, IEEE J. Biomed. Health Inform. 19 (6) (2014) 1893–1905.

[48] J. Chung, J. Teo, Mental health prediction using machine learning: Taxonomy, applications, and challenges, Appl. Computat. Intell. Soft Comput. 2022 (2022) 1–19.

[49] C. Molnar, Interpretable Machine Learning, Lulu. com, 2020.

[50] M.A. Ahmad, C. Eckert, A. Teredesai, Interpretable machine learning in healthcare, in: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 2018, pp. 559–560.

[51] J. Meisner, S. Rasmussen, M.E. Benros, Towards precision psychiatry utilizing large-scale multimodal data paving the way for improved prevention and treatment of mental disorders, Neurosci. Appl. (2022) 101017.

[52] M. Ferrara, G. Franchini, M. Funaro, M. Cutroni, B. Valier, T. Toffanin, L. Palagini, L. Zerbinati, F. Folesani, M.B. Murri, et al., Machine learning and non-affective psychosis: Identification, differential diagnosis, and treatment, Curr. Psychiat. Rep. (2022) 1–12.

[53] M. Ferrara, G. Franchini, M. Funaro, M. Belvederi Murri, T. Toffanin, L. Zerbinati, B. Valier, D. Ambrosio, F. Marconi, M. Cutroni, M. Basaldella, S. Seno, L. Grassi, Machine learning for mental health. Focus on affective and non-affective psychosis, in: Advancements in Artificial Intelligence in the Service Sector, Apple Academic Press, 2023.