SN Computer Science (2024) 5:30
https://doi.org/10.1007/s42979-023-02357-0

ORIGINAL RESEARCH

A Comparative Analysis on Recent Methods for Addressing Imbalance Classification

Zahid Ahmed[1] · Sufal Das[1]

Received: 12 June 2023 / Accepted: 22 September 2023

Abstract
In machine learning, the term ”class imbalanced” is frequently used. This is a crucial part of the field of machine learning. It is quite important in the classification process and has a significant impact on performance. That is why researchers are concentrating on it to overcome this difficulty. Various researchers have devised numerous methods till now. The approaches to addressing this imbalance issue found so far can be broadly categorized into three categories, which are the data-level approach, algorithm-level approach, and hybrid-level approach. To evaluate the most recent developments in resolving the negative effects of class imbalance, this study provides a comparative analysis of research that has been published within the last 5 years with an emphasis on high-class imbalance. In this study, an attempt has been made to provide a concise overview of what imbalance classification is, how it is created, and what the inconveniences are due to it. We have tried to provide a summary of several studies that have been published in the last few years and along with that a comparative analysis of all these approaches has been done.

Keywords Imbalanced classification · Majority class · Minority class · Data-level approach · Algorithm-level approach · Hybrid approach

Introduction

In today’s world, data are the most crucial aspect. It is fair to assume that data are essential for the advancement of science and technology, because data are required for the development of an automated system. In this current age of information exploration, the generation and collection of data are dramatically increasing, which results in a large amount of data [1]. Data play a key role in information processing, extraction, retrieval, and management [2].

Huge data are being generated every day and it is saved in databases. These data include machine-generated data, web, and social data, transaction data, human-generated data, biometrics data etc. [3, 4]. These data play a vital role in every automation system used in banking, healthcare, securities, education, communications, media, entertainment providers, and so on. During the data collection process, various sorts of data are acquired [3, 5–7]. Due to this, data handling has become a more challenging task. The largest issue in processing data is correctly classifying them and during classification, and imbalanced data are produced due to the disparity of instances in different classes. It is very essential to have proper classification of data to develop an accurate automation system. Because of the imbalance in training data, the rate of true negative and false positive increases. Many problems have to be faced due to imbalanced data, and as a result, it has become a popular topic among researchers. Many researchers have attempted to overcome the problem of imbalanced classification in recent years. In the last few years, a lot of new techniques have been proposed. In this study, we have attempted to incorporate some of them.

Imbalanced data are basically a classification problem that occurs when there are not an equal number of instances present in each classes [5, 8–16]. This means that some classes may have a very high number of instances, while

This article is part of the topical collection “SWOT to AI-embraced Communication Systems (SWOT-AI)” guest edited by Somnath Mukhopadhyay, Debashis De, Sunita Sarkar and Celia Shahnaz.

✉ Sufal Das
sufal.das@gmail.com

Zahid Ahmed
zahidprince786@gmail.com

[1] Department of Information Technology, North-Eastern Hill University, Shillong, Meghalaya 793022, India

Published online: 21 November 2023

others may have a very low number. If there are two classes of data, one class contains a small amount of data, while the other contains a very large amount.

In an imbalanced environment, the majority class is the class that has a large number of instances and the minority class is the one that has a very few number of instances.

In the process of data classification, the class distribution may be skewed for a variety of reasons. There are two main sets of grounds [5] for the imbalance that we should consider. The first one is the Sampling of data. It is frequently observed that there is an imbalance in classification at the time of sampling [5]. The reason for this can be either due to the errors encountered during the collection of data or erroneous measurement of data. For instance, possibly examples were taken from a limited geographical area or period, and the distribution of classes may be somewhat different or even collected in a different manner. One type of error is labeling multiple instances with incorrect class labels. Alternatively, the imbalance has been caused by damage or impairment to the processes or systems from which the examples were obtained [5]. The second one is the attributes of the domain. The imbalance may be a characteristic of the problem domain. For instance, one class' natural occurrence or presence may predominate over others. This could be because the procedure for generating observations in one class is more costly in terms of time, money, computing, or other resources. As a result, just collecting more instances from the domain to increase the class distribution is often impractical or impossible. To learn the differences between the classes, a model is required [5, 14–17].

Most classification algorithms perform well when the number of instances of each class is equal. When the number of instances of one class exceeds the number of instances of the other, problems arise. This is where the issue begins and it remains an outstanding problem. It generates a number of consequences like the number of minority class instances compared to majority class instances is extremely low [13, 14, 16, 18]. Due to the tiny size of the minority class, there is a lack of data available in the training data set [13, 14, 16, 18]. Small classes are often overlooked by classifiers, who instead focus on accurately classifying large ones[13, 14, 16, 18]. It is quite challenging to separate the minority class from the majority class [13, 14, 16, 18]. The majority of common classifiers presume that all domain application data sets are equal, although numerous data sets have a class imbalance distribution [13, 14, 16, 18].

The imbalance issue is a very serious matter due to which a lot of work has been done over the last few years. These approaches are based on three strategies, data level, algorithm, and hybrid-level strategy [19].

Data-level strategies or resampling techniques are very easy and cost-effective techniques to handle imbalanced data. Here, the distribution of the classes is balanced by either raising the minority class instances known as oversampling, or decreasing the majority class instances known as undersampling or by combining the two [13, 14, 16, 18, 20]. Random Undersampling [20] is a common undersampling strategy that randomly reduces the number of instances from the majority class, whereas synthetic minority oversampling technique (SMOTE) [21] is a common oversampling approach that generates synthetic instances in the minority class. These methods make it possible to train the model using a data set that is more evenly distributed [20]. For instance, SMOTE can be used to create fake fraudulent transactions for a more balanced data set in a credit card fraud detection system where actual fraud incidents are uncommon. Algorithm-level strategies concentrate on altering the machine learning algorithms to properly handle imbalanced data. To do so, the decision threshold of the algorithm may need to be changed. Sometimes cost-sensitive learning may be adopted, or ensemble techniques with built-in support for imbalanced data such as random forest or gradient boosting may be used [13, 14, 16, 18, 20]. For instance, altering the threshold for classifying an email as spam can give recall (accurately identifying spam) a higher priority than accuracy (avoidance of false positives). To build more accurate and fair models that are well-suited for imbalanced data, data and algorithm-level techniques are merged together which is known as a hybrid strategy. It utilizes the advantages of both data-level and algorithm-level techniques, offering a comprehensive solution to the imbalance classification problem [13, 14, 16, 18, 20]. For instance, in medical diagnosis, a hybrid approach may involve undersampling the majority class and subsequently training a support vector machine (SVM) algorithm with an adjusted decision threshold. This ensures that the model is sensitive to the detection of rare medical conditions.

We have done our study on a few of these approaches, among them. The following are the primary contributions of this study:

- This study will provide a basic overview of imbalanced data and the challenges related to it.
- This study will provide a general overview of the various strategies that can be used to balance data.
- A few of the recently proposed approaches are summarised as well as an attempt has been made to provide a comparative analysis of those approaches in this study.

The remaining of this paper is organized as follows:

For a better understanding of the current research techniques for solving the class imbalance problem, the details of some of the selected approaches proposed in the last 5 years are summarized in “Recent Works”. In “Discussion”, a comparative analysis of the aforementioned methodologies along with a brief discussion on the advancements in class imbalance research as well as how these trends and developments may affect future studies are summarized. Finally, our study ends with a conclusion in "Conclusion".

Recent Works

There are lots of approaches that have been proposed in recent years to address the problem of class imbalance. These approaches are based on any of the above-mentioned strategies.

Based on Data-Level Strategy

This type of strategy focuses on the data to achieve well-balanced classes. To balance the instances of different classes, numerous techniques are employed on data. It includes increasing instances of the minority class with respect to the instance of the majority class or decreasing the instances of the majority class with respect to the instance of the minority class. During this process, important instances may be eliminated or duplicate instances may be synthesized in an attempt to balance the classes [13–16]. Some of the approaches based on data-level strategy and the working principles are summarised one by one.

LoRAS: (An Oversampling Approach for Imbalanced Data Sets)

Localized random affine shadow sampling is an oversampling algorithm that has been proposed by Bej et al. [22]. The restrictions of SMOTE can be overcome with this approach. This approach is made to learn from a data set by roughly modeling the underlying instance manifold, assuming a set of features as best to utilize so that it can represent the data. It considers all features to be equally important. It synthesizes instances in the minority class using a parent data point from the minority class to get the k-nearest neighbor (k-NN) [23]. Next, it creates a shadow point for each data point in the neighborhood and then appends the shadow point to the neighborhood shadow sample. To construct the shadow samples and add noise to each feature, a list of standard deviations for normal distribution is created. The process of choosing a random shadow instance from a neighborhood shadow instance is then repeated until normalized random weights are produced for the selected spots. High dimensional data sets with more than 100 features and highly imbalanced data sets with an imbalance ratio greater than 25:1 were taken into consideration from the the Scikit-learn [24] library for the experiment.

Neighbourhood-Based Under-Sampling Approach for Handling Imbalanced and Overlapped Data

The NBUA is an under-sampling framework to eliminate potentially overlapping instances to handle a class imbalance in binary data sets. This approach was suggested by Vuttipittayamongkol and Elyan [25] and is intended to locate and get rid of excessive instances of a class in an overlapped region. The k-NN [23] rule is used to investigate each instance's local environment to reduce excessive deletion. By minimizing information loss and maximizing sensitivity, a close-to-ideal trade-off was made possible. This strategy relies on Four K-NN-based under-sampling techniques: Basic Neighbourhood Search (NB-Basic), Modified Tomek Link Search (NB-Tomek), Common Nearest Neighbours Search (NB Comm), and Recursive Search (NB-Rec). NB-Basic was developed to eliminate negative instances from the overlapped zone without considering any positive instances. This method could result in important data loss in accuracy, though, if negative instances are eliminated too quickly. Because of this, NB-Tomek and NB-Comm were developed to solve the issue of possible over-elimination of negative instances. Aiming to improve the detection of overlapped negative instances, NB-Comm was then extended to NB-Rec. For the experiment, 66 data sets with an IR of 1.86–41.4, 5–18 features, and 214–5472 instances were retrieved from the KEEL [26] and UCI [27] repositories. Both the handwritten digits data set from the MNIST database and the breast cancer data set from KDD Cup 2008 are also used.

Noise-Adaptive Synthetic Oversampling Technique

The noise-adaptive synthetic oversampling technology (NASOTECH) is a study that Vo et al. have suggested to address the class imbalance problem in imbalanced and noisy data sets [28]. First, the NASO technique was proposed to synthesize instances produced for each instance in the minority class, and it is based on the notion of the noise ratio. This methodology is the extended version of NASO. Here three variables are first initialized for the minority class, the majority class, and the balanced data set. The desired balancing ratio is then computed for the entire number of generated synthetic data samples. The k-NN of each sample in the minority class is identified. Following that, the total distance between each sample and K-NN samples is calculated. The number of generated synthetic data samples and each sample's noise ratio is determined. Finally, it creates synthetic samples for each sample and adds them to the balanced data set. The scene data set was collected from the LIBSVM Repository, and the remaining 9 are from the UCI Machine Learning Repository, with IR ranging from 28:1 to 8.6:1 for the experimental setup.

Neural Network-Based Under-Sampling Techniques

A neural network-based under-sampling algorithm has been proposed by Arefeen et al. [29]. Here, the minority class is trained using an auto-encoder and a straightforward ANN. After that based on either hard NN-based under-sampling or soft NN-based under-sampling strategy, it processes the minority instance. To determine which type of NN to be used to train the minority instance, a threshold value is established. To handle the minority instances properly either an auto-encoder is used if the number of input characteristics is more than the threshold value or a straightforward NN with two-thirds hidden layers is used if the number of input characteristics is less than the threshold value. In the end, a balanced data set is produced. For the experiment, 4 data sets from the UCI Machine Learning Repository, with IR varying from 1:78 to 11:8 are used.

Constrained Oversampling: (An Oversampling Approach to Reduce Noise Generation in Imbalanced Data Sets with Class Overlapping)

When data sets contain class overlapping regions, oversampling techniques introduce noise instances into the data sets. Liu et al. [30] have suggested a brand-new oversampling technique called Oversampling to lessen noise production. This approach executes in three steps. At first, the overlapping regions included in the original data set are extracted using a k-NN-based algorithm and eliminated from the overlapped area. Next, a boundary is established using an ant colony optimization algorithm. Here the borders between classes are retrieved and overlapped regions are defined. A collection of majority instances is used to represent the separation between majority and minority regions. Finally, synthetic instances are added. This stage involves building the final training set by synthesized instances from overlapping regions that are minorities and boundary instances. It initiates distance limitations on the oversampling process to lessen the development of fabricated minority instances in majority regions. It provides few restrictions to limit the additional instances to those areas where minority instances are already present but not to those areas where majority instances are present. For the experiment, five data sets from the UCI Repository were used.

Under-Sampling with Support Vectors for Multi-class Imbalanced Data Classification

Using a two-step under-sampling technique, Krawczyk et al. [31] have developed a novel method for handling multi-class unbalanced data. It extracts the core support vectors for each class using a one-class decomposition and uses these vectors as input prototypes for evolutionary under-sampling. In the first phase, a one-class classifier is trained on each of the classes to produce skew-insensitive data descriptions. It is possible to drastically minimize the number of instances required by extracting support vectors for each class and using them as new class representatives. The second phase involves using an evolutionary under-sampling technique to these support vectors to further balance the training set. Applying this method to a subset of support vectors rather than the entire data set decreases the computation time and increases accuracy. In the end, a balanced data set is ready to train a standard multi-class classifier. The methodology was tested on twenty multi-class unbalanced data sets that were taken from the UCI library.

SMOTE-IPF: Addressing Noisy and Borderline Examples in Imbalanced Classification

Saez et al. [32] have proposed an extended version of synthetic minority oversampling technique (SMOTE) by adding a new element iterative-partitioning filter (IPF). It is mostly observed that the performance of a model is affected by the noise and borderline instances. To increase the reliability and accuracy of imbalanced classification models, this approach tries to address the issue of noisy and borderline instances. It creates synthetic instances for the minority class, just like its predecessor SMOTE. However, it brings about a crucial improvement in the choice of "parent" samples. It considers the proximity of cases to make intelligent decisions as opposed to randomly choosing parent instances. By doing this, it ensured that the synthetic samples were produced in regions of the feature space where they would have a greater influence. The use of an intelligent filtering system makes it better than others. After creating artificial instances, it assesses how closely they resemble the dominant class. Noise introduction risk is decreased by filtering away instances that are too near to the majority class. The quality of synthetic samples is greatly enhanced by this filtering process. The trials were carried out on both synthetic and real-world data sets, with varying amounts of noise and morphologies of borderline samples. In addition, the influence of adding different types and levels of noise to these real-world data is investigated. They have used KEEL (Knowledge Extraction Based on Evolutionary Learning) data set.

Overlap-Based Undersampling for Improving Imbalanced Data Classification

Unlike other under sampling strategies based on clustering, Vuttipittayamongkol et al. [33] has proposed a framework that employs membership degrees obtained from the clustering process to aid in the removal of negative instances from the overlapping region. It starts with identifying instances in the majority class that overlap with the minority class using Fuzzy C-Means (FCM). Overlapping instances are crucial, as they result in ambiguity in classification. After the identification of overlapping instances, it applies a deterministic removal strategy. It selectively removes instances from the majority class that significantly overlap with minority class instances. It is designed to preserve the diversity of the majority class. Instead of arbitrarily eliminating instances, it carefully prunes those instances that cause overlap, ensuring that the majority class instances that remain are still representative. Data sets are obtained from UCI and KEEL repositories.

Based on Algorithm-Level Strategy

In this strategy, the existing learner is modified to remove its bias against classes. The most common approach is cost-sensitive learning, which forces the learner to correctly identify minority class instances by imposing a high penalty on incorrect minority class classifications. While there is no penalty for correctly classifying instances. It is observed that minority instances have a higher miss-classification cost than majority instances. The emphasis is given to minimizing the overall cost of the training data set. Since they depend on numerous circumstances, cost values are challenging to ascertain. References [13–15]. Here, we have attempted to examine a few of the most recent approaches that are based on algorithm-level strategy.

Fuzzy Support Vector Machine for Imbalanced Data with Borderline Noise

The Gaussian fuzzy function and a new distance metric have been proposed by Liu [34]. It has been developed based on the FSVM-CIL approach. The performance of any classifier is significantly impacted by the noise. Compared to other locations, the border region has maximum noise. First, noise in regions far from the borderline can be easily distinguished and managed, but noise near the borderline may be different from noise along the classification hyperplane. Using the distance measure and fuzzy function, borderline noise's impact on the FSVM-CIL can be lessened and performance can be improved. To accomplish this they have changed the existing distance measures and introduced a new one. A new fuzzy function is also introduced. On 25 publicly available imbalanced data sets from the KEEL data repository, experiments have been conducted.

A Novel Density-Based Adaptive K Nearest Neighbor Method for Dealing with the Overlapping Problem in Imbalanced Data Sets

The performance of the classifier is significantly impacted by the overlapping problem. That is why a density-based adaptive k-NN approach that can simultaneously manage imbalanced and overlapping issues has been suggested by Yuan et al. [35]. To proactively locate the most trustworthy query neighbors, they have created a straightforward and efficient distance adjustment technique. At first, using a density-based technique, training data are divided into six sections. After that, a distance metric is altered for each section by considering both local and global distribution. Finally, the query neighbors determined by the new distance measurements are used to create the output. This method changes the query neighbors based on how much imbalance and overlap are there. 41 data sets with having imbalance ratio ranging between 1.8 to 68.1 from the KEEL repository are used for practical purposes in the experiment. According to Fisher's discriminant ratio (F1), data sets are separated into two categories: low overlapping data sets with F1 greater than 1.6 and high overlapping data sets with F1 less than 1.6. Data sets are then sorted based on overlapping degrees.

Least Squares KNN-Based Weighted Multiclass Twin SVM

A weighted multi-class twin support vector machine based on least squares KNN has been presented by Tanveer et al. [36]. It is an expanded version of KWMTSVM. It can substitute equality constraints for inequality constraints and use the squared loss function as opposed to the hinge loss function used in Twin-KSVC and KWMTSVM. To take advantage of intra-class and inter-class information, the K-nearest neighbor graph technique is used, and different weight matrices are assigned to training data points for the same class. This method is incredibly easy and quick since it can solve two systems of linear equations instead of calculating QPPs in Twin-KSVC and KWMTSVM. It does not need a unique optimizer. This technique evaluates all the data instances into a "1-versus-1-versus-rest" structure, much like other SVM, to introduce ternary outputs that aid in handling imbalanced data sets. On 18 imbalanced data sets from the KEEL repository and the UCI machine learning repository, tests have been done.

A New Fuzzy K-Nearest Neighbor Classifier Based on the Bonferroni Mean

A novel approach has been put out by Kumbure et al. [37]. It is a generalized fuzzy k-nearest neighbor (FKNN) classifier that makes use of local mean vectors and the Bonferroni mean. The parametric Bonferroni means makes it possible to fit parameter values for a variety of situations and applications. It can function successfully even when there are huge imbalances in the data distributions. This technique generates local mean vectors for all classes that are represented by the k nearest neighbors using local sub-samples. Rather than directly comparing the query sample to the initial k nearest neighbors, it uses locally constructed representative vectors for each class that are well-positioned to perceive the class information. The use of local methods helps to solve the issues of class imbalance. Furthermore, issues that arise when using imprecise data in scenarios where samples from various classes are extremely near to one another can also be fixed. Accurate classification usually depends heavily on the choice of the k value. A very low k value can lead to inaccurate classification findings, whereas a large k can lead to outliers having an impact on the classification. The k values chosen in the context of the proposed method can be relatively high, allowing the method to capture bigger class-representative sub-samples and provide more precise local Bonferroni mean vectors. In the experiment, six real-world data sets from KEEL and the UCI Machine Learning repository are utilized.

Deep Reinforcement Learning for Imbalanced Classification

When the distribution of the data is uneven, conventional classification techniques are ineffective and may even fail. A universal imbalanced classification model based on deep reinforcement learning has been suggested by Lin et al. [38] to address this problem. This method formulates the classification problem as a series of sequential decisions and then uses a deep Q-learning network to solve it. At every stage, the agent classifies one instance, and the environment assesses the classification action and gives the agent a reward. Because the reward from the minority class instance is higher, the agent is more aware of the minority class. Under the direction of a particular reward function and a helpful learning environment, the agent eventually discovers an ideal classification policy in imbalanced data. They have used the IMDB, Cifar-10, Mnist, and Fashion-Minist data sets to conduct the tests.

Affinity and Class Probability-Based Fuzzy Support Vector Machine for Imbalanced Data Sets

When dealing with classification issues imposed by an imbalanced data set, a conventional SVM can typically demonstrate reasonably robust performance. However, because it treats all training instances equally when learning, the final decision boundary will skew toward the majority class, especially when outliers or noise are present in the data set. A novel affinity and class probability-based fuzzy support vector machine approach has been proposed by Tao et al. [39] A support vector description domain model, similar to the one used for FSVM learning, is used to determine the affinity of a majority class instance using only the majority class training examples that were provided. To uncover some border samples and potential outliers, the affinity that was obtained can be applied to the majority class training data. To reduce the effects of noise, the kernel k-nearest neighbor method is used to calculate the class probability of the majority of class instances in the same kernel space. Low memberships that may be deduced from the class probabilities and affinities appear to limit the learning capacity of the instances. Lower class probabilities show that the instances are more likely to be noisy. The final classification border is shifted toward the majority class. As a result, it assigned less weight to instances from the minority class with lower affinities and class probabilities and more weight to instances from the majority class with higher affinities and class probabilities. Relatively high memberships are also assigned to the minority class instance to ensure their significance for the model learning. For the experiment, 27 different data sets were chosen from the UCI Machine Learning Repository.

Based on Hybrid Strategy

These strategies are a mix of both data-level and algorithm-level strategies outlined above. The main goal of these is to improve prediction performance when compared to using only one classifier. The fundamental issue with these methods is that they generate more classifiers, which increases the computing complexity. References [13–15, 40]. We have made an effort to study some of the recently proposed approaches based on a hybrid strategy.

A Weighted Hybrid Ensemble Method for Classifying Imbalanced Data

For categorizing imbalanced data in binary classification, Zhao et al. [41] have suggested a weighted hybrid ensemble technique. The proposed method, which fits into the boosting algorithm's framework, combines two data sampling techniques with two base classifiers, and each sampling technique and each base classifier receive matching weights to improve them. Here, random undersampling and adjustable random balance are used as sampling techniques. Support vector machines and decision tree classifiers are used as basic classifiers. A scale factor is used to regulate the range in which the number of class instances can fluctuate and can prevent a class from having an excessively small or excessively large number of instances. Using this scale factor, it can also guarantee that each class keeps a specific amount of instances, producing outcomes that are superior to the initial random balance. For the experiment, 31 data sets from the KEEL data repository with imbalance rates ranging from 1.87 to 129.44 and 9 data sets from HDDT with imbalance rates ranging from 2.41 to 42 are employed.

Ensembling Perturbation-Based Oversamplers for Imbalanced Data Sets

Zhang et al. [42] have proposed an approach that initially trains a large number of classifiers from balanced subsets generated by the perturbation-based oversampling (POS) approach, and then uses majority voting to fuse them into an ensemble. As a result, the suggested method is called the perturbation-based oversampling ensemble. How to produce varied subsets for classifier training is a critical aspect of ensuring excellent ensemble learning performance. There are two variants of POSENS have been introduced. The first is to use the random subspace approach to identify which features should be disrupted in subsets. The second is due to the POS's nature, which generates new instances by randomly perturbing features of previously generated seed instances. To avoid creating instances in inappropriate regions, the Bayes' rule is used to compute the sensitivities of minority instances concerning class imbalance, and new examples are synthesized based on those with high sensitivities. These strategies ensure that the POSENS generates a diversified range of subsets and that the final classifier performs well. Thirty-five imbalanced data sets to evaluate the performance of the proposed method are used from the UCI and the KEEL data repositories.

Hybrid Neural Network with Cost-Sensitive Support Vector Machine for Class-Imbalanced Multimodal Data

To deal with class-imbalanced in multi-modal data, Kim et al. [43] have presented a hybrid neural network with a cost-sensitive SVM. They have used cost-sensitive support vector machines as a classifier and a fused multiple-network structure constructed by separating the features from data from several modalities. To manage heterogeneous data, feature extraction is done first in the fusion NN architecture. The NN architecture consists of multilayer perceptrons for extracting information from structured numeric data and convolutional NNs for natural language processing (NLP). The retrieved features are supplied into the CSSVM classification layer, which uses a novel gradient-descent method to minimize the loss function produced from an L2-CSSVM to fine-tune the entire model, to improve classification performance in a class-imbalanced situation. To conduct the trial, real-world imbalanced data sets from the KEEL repositories like Wine Review, Yelp, etc. are used.

A Hybrid Classifier Combining SMOTE with PSO to Estimate 5-Year Survivability of Breast Cancer Patients

Wang et al. [44] have proposed a hybrid approach that combines Particle Swarm Optimisation (PSO) and the Synthetic Minority Over-sampling Technique (SMOTE) with logistic regression, the C5 decision tree (C5) model, and a 1-nearest neighbor search. The class imbalance in the data set is addressed using SMOTE. To improve the minority class's representation, it creates fake samples for them. The feature selection process uses PSO. The subset of features that are most useful for prediction are optimized through PSO. PSO adjusts the classifier to the specific features of the data set by experimenting with feature combinations. The output of the SMOTE-augmented data set and the PSO-selected feature subset is fed into the classifier, which exploits the advantages of both approaches and combines the upgraded data set and the optimized feature subset for improved classification performance. The data set of breast cancer patients from SEER was used to categorize patients based on their 5-year survival rates. When combined with the right searching algorithms like PSO and classifiers, SMOTE can significantly increase the effectiveness of classification for hugely unbalanced data sets.

Discussion

This study is done on six data-level approaches, six algorithm-level approaches, and three hybrid approaches. To evaluate the performance, effectiveness, and efficiency of a system, evaluation metrics are used. These are the quantitative and objective criteria or characteristics of a system. These are also referred to as indications or performance measures [45]. After studying all the approaches, it has been observed that the evolution metrics used to test out every approach are any of the following evaluation metrics discussed below.

For a binary data set having positive P and negative N instances are classified. By considering a threshold value, if a rank of probability is calculated for each P and N, in which class it will fall? An instance having a rank greater than the threshold value as a positive class is considered positive. That means if a P is positive it is true positive (TP) [45] else it is false positive (FP) [45]. Similarly, if a N is negative it is true negative (TN) [45] else it is false negative (FN) [45]. The proportion of genuine positive cases that are accurately recognized by a classification model is measured by the true positive rate (TPR), which is also known as sensitivity or recall [45]. The proportion of genuine negative instances that are correctly classified as negative by a classification model is measured by the true negative rate (TNR), which is also known as specificity [45]. The proportion of genuine negative instances that a classification model wrongly classifies as positive is measured by the false positive rate (FPR) [45]. The proportion of genuine positive instances that a classification model wrongly classifies as negative is measured by the false negative rate (FNR), which is also referred to as the miss rate [45]. A high TPR indicates that the model has a low percentage of miss classification and is an effective one, a high TNR indicates that the model has a low rate of miss classification and is effective at finding negative instances, a high FPR indicates that the model frequently wrongly identifies negative instances as positive and a high FNR indicates that the model is more likely to miss classify positive instances as negative instance [45]. Area under the curve (AUC) offers a single scalar value that measures a model's overall capacity to discriminate between positive and negative classifications. The area under the receiver operating characteristic (ROC) curve is measured by AUC. The ROC curve visually illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (specificity) at various thresholds. Recall is the percentage of accurate positive predictions made out of all accurate forecasts, and Precision is the percentage of accurate positive predictions made out of all instances where a prediction was correct. Precision and Recall are balanced by the F-measure also commonly referred to as the F1-score. The degree to which a model successfully predicts the classes of an instance is known as Accuracy and the degree to which a model wrongly classifies classes of an instance is known as inaccuracy [45]. Geometric mean (G-Mean), is used to assess how well a binary classification model performs. It is used to determine a model's capacity to produce highly accurate results for both the positive and negative classes at the same time [45]. An average accuracy (AvAcc) provides an indication of how consistently a model performs well in various scenarios. Cost–benefit analysis (CBA) is a technique for determining whether using a specific model is cost-effective or not [45].

To compare the performance of all approaches either the best approaches among the existing are used as the baseline algorithm, or the approach that has been proposed earlier is used as a baseline. The formula for the evolution metrics discussed above is tabulated in Table 1.

Table 1 Evolution metrics formulas

Sl. no.	Metrics	Formula
1	TPR [45]	$\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$
2	TNR [45]	$\text{TNR} = \frac{\text{TN}}{\text{TN+FP}}$
3	FPR [45]	$\text{FPR} = \frac{\text{FP}}{\text{FP+TN}}$
4	FNR [45]	$\text{FNR} = \frac{\text{FN}}{\text{FP+TP}}$
5	Sensitivity [45]	$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}}$
6	Recall [45]	$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$
7	Specificity [45]	$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}}$
8	Precision [45]	$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$
9	AUC [45]	$\text{AUC} = \frac{1+\text{TPR}-\text{FPR}}{2}$
10	F-measure [45]	$\text{F-measure} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}}$
11	Accuracy [45]	$\text{Accuracy} = \frac{(\text{TP+TN})\times 100}{\text{TP+FP+TN+FN}}$
12	Inaccuracy [45]	$\text{Inaccuracy} = \frac{(\text{FP+FN})\times 100}{\text{TP+FP+TN+FN}}$
13	G-Mean [45]	$\text{G-mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}}$
14	AvAcc [45]	$\text{AvAcc} = \frac{\sum_{i=1}^{M} TPR_i}{M}$
15	CBA [45]	$\text{CBA} = \frac{\sum_{i=1}^{M} \frac{mat_{i,j}}{max(\sum_{i=1}^{M} mat_{i,j},\sum_{i=1}^{M} mat_{j,i})}}{M}$

Data-level perspective: Four under-sampling techniques—NBUA, NNBUT, USVMIDC, and OBUS as well as four over-sampling techniques—LoRAS, NASOTECH, COA, and SMOTE-IPF were taken into account for this study.

LoRAS is an oversampling technique. Typically, LoRAS results in less misclassification for the majority class with just a modest amount of room for error in the minority class. This approach can estimate more accurately the mean of the underlying local distribution for a minority class sample for tabular high-dimensional and highly imbalanced data sets. It had limitations for imbalanced data sets based on diverse images. This method outperforms synthetic minority oversampling technique (SMOTE) [21] extensions like adaptive synthetic sampling (ADASYN) [46], support vector machines–SMOTE (SVM–SMOTE) [47], Borderline1 SMOTE (B-SMOTE) [48], and Borderline2 SMOTE [49]. This approach is developed as an alternative to SMOTE for processing extremely imbalanced data sets [22].

When there is an overlap between instances from distinct classes, the learning task becomes more difficult. The NBUA under-sampling framework eliminates potentially overlapped instances to handle a class imbalance in binary data sets. It accurately locates and removes majority class instances from the overlapped region. Additionally, it can limit data loss caused by excessive data deletion. In comparison with class distribution-based approaches like SMOTE, k-means under-sampling, and class overlap-based methods like overlapped-based undersampling (OBU), Borderline-SMOTE (BLSMOTE) [48], edited nearest neighbour (ENN), support vector machine (SVM), and random forest (RF), it produces better results. Eventually, it is not appropriate for both high-density base data and data from the actual world. It is also limited to binary class only [25].

Most current oversampling techniques are lacking a method for dealing with noise instances in imbalanced and noisy data sets, which lowers the predicted accuracy of machine learning models. NASOTECH is capable of addressing the issue of class imbalance in imbalanced and noisy data sets. Performance may go up as a result. In comparison to SMOTE, Borderline-SMOTE, and ADASYN, it can perform better. The drawback of this strategy is that it cannot effectively address the issue of class imbalance across a variety of fields. Certain optimization techniques could be added to enhance NASOTECH's performance [28].

NNBUT is an NN-based under-sampling strategy. NN has the ability to identify complex patterns in instances to address the problem of class imbalance. Less overlapping in the instance improves the performance of hard NN-based under-sampling. The majority instances, which are far from the minority instances, are actually retained. Soft NN-based under-sampling works better in cases with overlapped data. This strategy is tested using different classifiers, and the performance results show a significant improvement over ENN, all k-nearest neighbors editing (AKNN), Near Miss method-1 (NM-1), NM-2, NM-3, neighborhood cleaning (NCL), Random Undersampling (RUS), and Tokem Link (TL). The key disadvantage of this strategy is the significant likelihood of information loss if the IR is very high [29].

With the addition of limitations to the oversampling process, COA stands apart from other oversampling methods by preventing noise creation in overlapping regions. It is more resistant to noise in the original minority set since it does not rely only on the information provided by minority instances. By locating the boundary majority of instances and including them in the oversampling procedure, it is desirable to widen the decision region for the minority category. It should be mentioned that only the impact of class overlapping has been evaluated using this strategy to address the imbalanced classification problem. There is still much work to be done to fully understand the impact of other instances in data sets that are imbalanced. According to experimental findings, this technique is superior to SMOTE, y Constrained Oversampling (CO), BOSMOTE-1, BOSMOTE-2, ADASYN, and Cluster-Based Synthetic Oversampling (CBOS). In spite of all these benefits, this technique has a high cost for calculation and storage [30].

For multi-class imbalanced data, USVMIDC is a powerful under-sampling approach that can perform better than over-sampling techniques. The majority of currently used strategies for addressing imbalanced data concentrate on oversampling techniques. In binary classes, it can alleviate several oversampling restrictions including increasing class overlaps, improving noise present, or modifying class distributions. This method can fill the gap left by under-sampling approaches that can assume inter-dependencies between classes and account for multi-class imbalance. This strategy performs superior to STATIC-SMOTE (S-SMOTE), Mahalanobis Distance Oversampling (MBO), (k-NN)-based synthetic minority oversampling (SMOM), Multiclass Evolutionary Undersampling (MC-EUS), and One-Class Support Vector base Undersampling (OCSV-US) in comparison. Instance-level and class-level challenges during the selection process cannot be handled by this strategy [31].

Although SMOTE generates a better distribution of examples throughout the classes, it has several downsides, such as the generation of an excessive amount of instances centered on pointless positive examples that do not aid minority classes in learning. The erasing of the boundaries between classes and the introduction of noisy positive examples in areas controlled by the majority class both result in an increase in class overlap. SMOTE-IPF is capable of resolving these issues. It is appropriate for imbalanced data sets with noisy and ambiguous examples. Iteratively removing noisy examples is possible. It can forecast more accurate noisy cases. Comparatively, it is a better approach than B1-SMOTE, B2-SMOTE, SL-SMOTE, SMOTEENN and SMOTE-TL. The biggest drawback of this process is the choice of various IPF parameters. As the behavior of the filter depends on the parameters and there are many parameters accessible, the various parameters can affect performance [32].

OBUS lessens the prevalence of occurrences from the majority class. It has the ability to recognize overlapped areas and remove any problematic occurrences from those areas. It can lessen misunderstanding and increase the learner's awareness of the positive examples. Additionally, it can lessen information loss. It is a better approach than K-means. It is a fairly slow process [33].

At the data level, either oversampling or under-sampling is employed to balance a data set. It has been established that under-sampling is preferable to over-sampling. This is due to the possibility of over-fitting during the model generation process being increased by the over-sampling method. Some useful data existing in the majority class can be lost if the under-sampling technique is used. The majority of these efforts are made to address these issues.

An attempt has been made to make a comparative analysis of all these data-level approaches in Table 2.

Accuracy can sometimes be misleading, especially in data sets with an imbalance where one class is vastly outnumbered. In situations where classes are imbalanced, the F1 score offers a more accurate picture of a model's performance. Let's take a closer look at each of these approaches according to their accuracy and f1 score. Considering the accuracy and f-1 score of all the eight data-level approaches discussed so far, which are evaluated using an SVM classifier on two different data sets namely Abalone and Yeast taken from UCI [27] data repository and are listed in Table 3.

From Fig. 1, it is clear that on the Abalone data set, LoRAS has an accuracy of 0.67 and an F1-score of 0.59, compared to NBU's accuracy of 0.52 and F1-score of 0.22. This shows that LoRAS outperforms NBU on this data set in terms of accuracy and F1 score. When choosing amongst these methods, it's crucial to take the unique problem into account as well as the trade-offs between recall and precision. On the Yeast data set, NASOTECH and NNBUT both display excellent accuracy and F1-score values, demonstrating strong performance.

This shows that these techniques work well for this specific data set. When choosing a strategy, it's necessary to consider additional elements like computational complexity and practical applicability. On both data sets, COA and

Table 2 Comparison table for data-level approaches

Approaches	Data sets	Evaluation metrics	Baseline algorithm	Findings	Limitations
LoRAS [22]	27 data sets from the theScikit-learn library [50] having IR is more than 25:1 and features more than 100	Accuracy, Sensitivity, Recall, F1-Score, and BA	SMOTE ADASYN SVM–SMOTE. Borderline 1–2 SMOTE	It can reduce misclassifications Better than SMOTE and its other extensions It can process extreme IR	Limited to data sets based on diverse images Not suitable for extremely large data sets Risk of over-fitting
NBUA [25]	Handwritten digits data set from the MNIST. Breast cancer data set from KDD Cup 2008 [51]. 66 Data set from the KEEL [52] and UCI [27] repositories with IR of 1.86–41.4, 5–18 features, and 214–5472 instances	Specificity Sensitivity G-mean F1-Score and Precision	SMOTE k-means OBU BLSMOTE ENN SVM RF	It can identify and eliminate overlapped data Reduce information loss It can minimize excessive elimination of data	Limited to binary class data sets Not suitable for high-density data sets Not applicable for real-world problems
NASOTECH[28]	Scene data set from the LIB-SVM Repository [53] and 9 data sets from the UCI [27] Machine Learning Repository, with IR 28:1–8.6:1	Accuracy. Specificity Sensitivity G-mean	SMOTE Borderline-SMOTE ADASYN	It can identify and eliminate noisy data It can increase performance It can perform better in low-level noise data set	Limited to Specific domain Not suitable in extremely noisy data sets Risk of over-fitting Addition of Certain optimization techniques can increase the performance
NNBUT [29]	4 data sets from the UCI [27] Machine Learning Repository, with IR 1:78 to 11:8	AUC ROC Precision Recall F1-Score	ENN AKNN NM-1-2-3 NCL RUS TL	Can generate the balance data set Average approach to deal overlapped data Suitable for low IR	Risk of information loss Not suitable for high IR Not an efficient approach
COA [30]	5 Data setsfrom UCI[27] repositories	G-mean.F-measure.overallaccuracy	SMOTE.COBOSMOTE-1-2. ADASYN.CBOS	It can preventnoisecreation inthe overlappedregion. It does notrely only onthe minoritysample'sinformation.It can locateboundaryregions	High computationalcost.High spacecomplexity.Not efficient
USVMIDC [31]	20 multiclassimbalanceddata setsfrom theUCI [27]repository	Averageaccuracy.Classbalanceaccuracy. Confusionentropy.G-mean	S-SMOTE.MBO.SMOM.MC-EUSOCSV-US	Strong atdeleting usefulexamplesfrom thetraining set. Redundancyin prototypescan bereduced.It can reducenoise fromthe data set	Not efficient.Can nothandle-Instancelevelchallenges.Can nothandleInstancelevelchallenges
SMOTE–IPF [32]	Nine noisyand borderlinereal-worlddata setsfromKEEL [52] repository	AUC	B1, B2-SMOTESL-SMOTES-MOTEENNSMOTE-TL	Capable ofeliminatingnoise. Handleborderlineinstances. ImprovedGeneralization	Weak parameters Computationalcomplexity. Dependenton datasetquality
OBUS [33]	Thirty-sixdata-setswith IR1.87 to129.44 fromUCI [27] and-KEEL [51]repositories	Accuracy Sensitivity	K-means	Eliminateinstancesoutside theoverlappingregion.Improvedrobustness	Relativelyslow.Risk of informationloss.Sensitivityto parameter-tuning

Table 3 Accuracy and F1 score using SVM classifier

Approaches	Data set	Accuracy	F1 Score
LoRAS [22]	Abalone	0.67	0.59
LoRAS [22]	Yeast	0.84	0.35
NBUA [25]	Abalone	0.52	0.22
NBUA [25]	Yeast	0.73	0.34
NASOTECH [28]	Abalone	0.7	0.74
NASOTECH [28]	Yeast	0.87	0.8
NNBUT [29]	Abalone	0.91	0.82
NNBUT [29]	Yeast	0.92	0.84
COA [30]	Abalone	0.93	0.8
COA [30]	Yeast	0.95	0.81
USVMID [31]	Abalone	0.84	0.86
USVMID [31]	Yeast	0.82	0.84
SMOTE-IPF [32]	Abalone	0.82	0.81
SMOTE-IPF [32]	Yeast	0.88	0.79
OBUS [33]	Abalone	0.55	0.69
OBUS [33]	Yeast	0.71	0.82

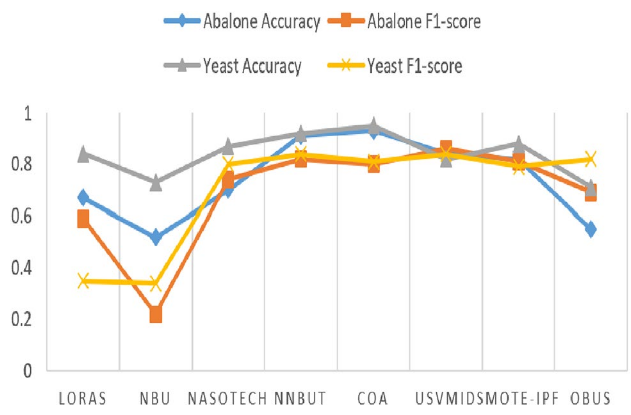

Fig. 1 Performance measured based on accuracy and F1 score

USVMID regularly outperform other models in terms of accuracy and F1 score. This shows that these methods may be applicable to a variety of imbalanced classification issues and may be robust across various data sets. SMOTE-IPF performs well in terms of F1-score, notably on the Yeast data set, while OBUS performs well in terms of accuracy and F1-score on both data sets. These findings show that SMOTE-IPF and OBUS are competing methodologies for the classification of imbalances.

Algorithm-level perspective: There are six algorithmic-level approaches considered in this study, those are FSVMIBN, DBANN, LS-KWMTSVM, BM-FKNN, DQN-imb, and ACFSVM.

The FSVM–CIL is extended by FSVMIBN. To increase the effectiveness, Borderline noise needs to be reduced, but FSVM–CIL neglected this issue. It is very important to reduce the effect of borderline noise. This strategy can able to reduce the noise at the borderline. It can deliver superior outcomes compared to the benchmark approaches of cost-sensitive SVM (CS-SVM), SMOTE–SVM, and SVM, respectively. The parameters that were taken into account in this approach are not sufficient enough which is why adding more parameters may improve the outcomes [34].

The idea of DBANN is to use density-based approaches to identify the most trustworthy query neighbors. In terms of average rank in F1 and GM, it outperforms alternative approaches in almost all data sets. The best average rank is obtained when the data distribution is highly overlapping. In all data sets, with the exception of a few, it yields the best results when the degree of overlap is modest or slight. For the imbalance problem, it performs well in high imbalance ratios while performing moderately in low imbalance ratios. It performs better than other kNN-based techniques such as Weighted k-NN (W-kNN), k Rare-class Nearest Neighbour (kRNN), k-NN, Fuzzy k-NN (F-kNN), k-local hyperplane distance nearest neighbor (H-kNN), and also Classification and Regression Tree+SMOTE (CART+SMOTE), CART+OBU, SVM+SMOTE, SVM+OBU, etc. It is limited to binary classification problems only [35].

LS-KWMTSVM produces an incredibly straightforward and quick algorithm by solving two systems of linear equations. As a result, it does not require an external optimizer like Twin-KSVC or KWMTSVM. To simplify the complex nonlinear LS-KWMTSVM, the Sherman–Morrison–Woodbury (SMW) formulation is used. It is able to evaluate all of the training data points in a "1-versus-1 and 1-versus-rest" framework, enabling it to produce ternary outputs of 1, 0, and 1, which aid in handling imbalanced data sets. To exploit intra-class and inter-class information, it employs a KNN graph technique, where data points inside the same class are assigned various weight matrices. It can outperform Twin Multi-class Classification Support Vector Machines (Twin KSVC), Least Squares Twin Multi-class Classification Support Vector Machine (LST–KSVC), and KNN-based Weighted Multiclass Twin Support Vector Machines (KWMTSVM) techniques in terms of performance. It is a practical and successful method for addressing imbalance classification issues. For situations involving imbalance classification, it is a successful and effective strategy. This approach has a significant problem with parameter selection, because there are many parameters available [36].

BM-FKNN performs better than FKNN, LM-KNN, KNN, and SVM. As opposed to membership degree computation in fuzzy k-NN, the influence of the Bonferroni means inside the learning portion of the classifier has a dominant effect. One can create more logical class representative vectors using the concepts of Bonferroni mean local vectors rather than k nearest samples as nearest representations. Even while SVM, NB, and similarity classifiers can sometimes attain somewhat higher accuracy levels, Mean-based Fuzzy k-Nearest Neighbor (BM-FKNN) and Mean-Based k-Nearest Neighbor (BM-KNN) classifiers still outperform them on most data sets. Additionally, it appears that using the higher values for the parameter k using this approach, the performance of the classifier has been greatly improved. The execution of this method takes a little longer time due to the requirement of more complicated computation. Furthermore, because of the utilization of grid search to obtain appropriate Bonferroni mean parameters usually it takes a little more time compared to traditional methods. It also has a relatively high computational complexity also [37].

The DQNimb technique frames the classification problem as a sequential decision-making process, where the environment delivers a large reward for minority class samples but a low reward for majority class samples and the process will end when the agent incorrectly classifies the minority class instances. The best classification strategy for the Imbalanced Classification Markov Decision Process (ICMDP) is determined using the deep Q learning algorithm, and theoretical analysis is done to determine how a certain reward function will affect the Q network's loss function during training. Reducing the reward value the agent receives from the majority of samples can balance the impact of the two types of samples on the loss function. Compared to other imbalanced classification techniques like e Deep Neural Network (DNN), this method performs better. It performs better, particularly with text data sets and highly imbalanced data sets, and it exclusively addresses the binary class imbalance issues only [38].

In the ACFSVM, the SVDD model is initially trained using the provided majority of instances. The proper formulation of affinity with relation to the trained SVDD model is then delivered to calculate a unique affinity for each instance in the majority class. Using this method. It is possible to effectively locate the border instances and probable outliers that exist in the majority class. To lessen the effect of class noises in the majority class, the membership value for a fuzzy SVM is calculated using the kernel k-NN technique, which first calculates the class probability for each instance belonging to a majority class along with its associated affinity. This method typically assigns relatively low MVs to some potential abnormal majority samples based on their corresponding affinities and class probabilities while the high MVs to rare minority instances, which can allow the final classification boundary to skew toward the majority class and produce more satisfying classification results. It can outperform Support Vector Machine (SVM), Cost-sensitive SVM (CSVM), Harmonic Element SVM (HesSVM), Random Undersampling-SVM (RUSVM), Random Oversampling-SVM (ROSVM), SMOTESVM, BSMOTESVM, WKSMOTESVM, AdaSVM, General Membership Function-SVM (GPFSVM), Entropy-based Fuzzy-SVM (EFSVM), Entropy-based Fuzzy Least Squares-SVM (EFLSSVM-CIL), and Entropy-based Fuzzy Least Squares Twin-SVM (EFLSTWSVM-CIL). If the ideal parameters can be identified, this strategy will perform better [39].

Table 4 will provide a comparative analysis of all these algorithm-level strategies discussed above.

An important performance metric in machine learning, especially for binary classification problems, is the AUC (Area Under the Receiver Operating Characteristic Curve). In contrast to the F1 Score, which evaluates the model's precision and recall in relation to each other, AUC evaluates the model's capacity to distinguish between positive and negative classifications. Let's take a closer look at each of these above-mentioned approaches based on their AUC and f1 score. Considering the AUC and F1 score evaluated on two different data sets namely Ecoli and Yeast taken from UCI[27] data repository and are listed in Table 5.

Figure 2 clearly shows that on both data sets, the FSVMIBN and DQNimb techniques provide excellent performance. Their AUC and F1 Score values are comparatively high. DQNimb stands out with nearly flawless AUC values of 0.95 on Ecoli and 0.97 on yeast. These findings imply that FSVMIBN and DQNimb are viable options for these data sets, because they balance precision-recall balance (F1 Score) and class separation (AUC) in a promising manner.

On both data sets, BM-FKNN performs favourably in terms of AUC and F1 Score. High AUC values are present, which suggests that it has good discrimination skills. Additionally, its F1 Score values are solid, showing that it balances the trade-off between recall and precision. According to this, BM-FKNN is a trustworthy method for binary classification tasks on yeast and Ecoli. AUC is 0.9 and the F1 Score is 0.91 for LS-KWMTSVM on Ecoli, which is a respectable performance. As the Yeast data set's AUC and F1 Score are comparatively smaller. This suggests that because the Yeast data set can have different properties, the technique might not be as appropriate for it. In comparison to the others, the performance of the DBANN and ACFSVM techniques is comparatively lower. The AUC values for DBANN on Ecoli and yeast are 0.71 and 0.75, respectively, indicating that it might have trouble with class separation. With an F1 Score of 0.79 on Yeast, ACFSVM likewise has issues. It is crucial to take these lower values into account in relation to the requirements of the particular challenge.

Hybrid-level perspective: This study considered four hybrid-level approaches to conduct a small investigation into them. These approaches are WHMBoost, POSENS, NN-CSSVM and SMOTE+PSO.

The distribution of data in actual data sets is typically imbalanced. The dominant class is often over-represented by the classifier, while the under-represented minority class is difficult to classify accurately. However, in real-world situations, everyone is typically more interested in the minority

Table 4 Comparison table for algorithm-level approaches

Approaches	Data sets	Evaluation metrics	Baseline algorithm	Findings	Limitations
FSVMIBN [35]	25 data sets from KEEL [52] repository	AUC	CS-SVM SMOTE–SVM SVM	It can detect borderline noise It can reduce borderline noise	Weak parameter. Performance can be increased by adding more parameters
DBANN [35]	41 data sets from the KEEL [52] repository with an imbalance ratio ranging between 1.8 to 68.1	G-mean F1-Score	W-kNN kRNN. kNN F-kNN. H-kNN	It can identify the most trustworthy query neighbors It can handle overlapping data Suitable for data sets with high IR	Limited to binary class data sets Not suitable for low overlapped data sets Not applicable for data sets with low IR
LS-KWMTSVM [36]	18 data sets from the KEEL [52] and UCI [27] machine learning repositories	Accuracy Execution Time	Twin-KSVC LST-KSVC KWMTSVM	It can eliminate redundant constraints It is relatively faster than the previous version It can solve a system of linear equations instead of QPPs	Many parameters are available Parameter selection is tough
BM-FKNN [37]	6 data sets from KEEL [52] and UCI [27] machine learning repositories	Accuracy Specificity Sensitivity	FKNN LM-KNN KNN SVM	Better classification accuracy Less sensitive to class imbalance	Relatively slow High computational complexity Not efficient
DQNimb [38]	IMDB, Cifar-10, Mnist, and Fashion-Minist data sets [54]	G-mean F-measure	DNN	Suitable for data sets with high IR	Limited to binary class data sets Limited to text data sets only Not Suitable for data sets with low IR
ACFSVM [39]	27 data sets from UCI [27] machine learning repository	G-Mean. F-Measure AUC	SVM and its extensions	It can identify the possible outliers and border samples existing in the majority class It can avoid the effect of class noises in the majority class	Weak Parameter Not efficient

Table 5 AUC and F1 score of the algorithm-level approaches

Approaches	Data set	AUC	F1 score
FSVMIBN [35]	Ecoli	0.96	0.91
FSVMIBN [35]	Yeast	0.87	0.82
DBANN [35]	Ecoli	0.71	0.69
DBANN [35]	Yeast	0.75	0.5
LS-KWMTSVM [36]	Ecoli	0.9	0.91
LS-KWMTSVM [36]	Yeast	0.78	0.75
BM-FKNN [37]	Ecoli	0.85	0.84
BM-FKNN [37]	Yeast	0.87	0.88
DQNimb [38]	Ecoli	0.95	0.86
DQNimb [38]	Yeast	0.97	0.98
ACFSVM [39]	Ecoli	0.83	0.86
ACFSVM [39]	Yeast	0.82	0.79

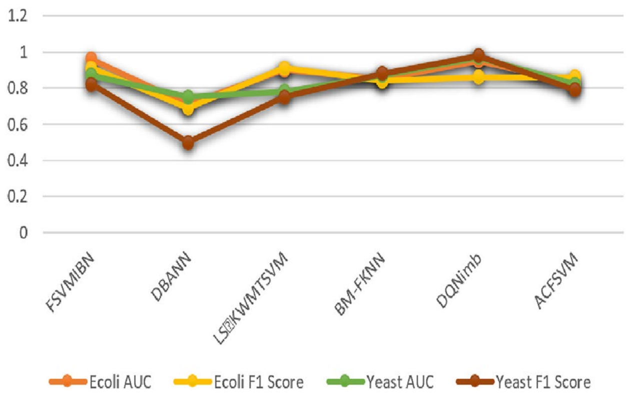

Fig. 2 Performance measured based on AUC and F1 score

class. In this case, it is quite difficult to classify the minority class accurately. WHMBoost can minimize the effects of data imbalance and increase the likelihood that occurrences of the minority class will be correctly classified by the model. Compared to employing a single sampling technique and a single basis classifier, it combines the benefits of multiple sampling methods and multiple base classifiers and enhances overall performance. Results of the experiments indicate that it is superior to Adaptive Boosting (Adaboost), Random Undersampling Boosting (RUSBoost), Random Balance Boosting (RBBoost), Random Hybrid Sampling Boosting (RHSBoost), SMOTE Boosting (SMOTEBoost), cluster-based undersampling Boosing (CUSBoost), and Multiple Estimators Boosting (MEBoost) [41].

By randomly perturbing random subsets of the input attributes of seed instances that have been provided, POSENS creates a variety of new instances. Sensitivity with respect to class imbalance is computed for each minority instance to reduce the possibility of noise introduction, and cases producing high sensitivities are more likely to be chosen as seed instances. These different subsets are utilized to train various base classifiers after producing new instances, which are subsequently fused to improve overall performance via a majority voting. The POSENS greatly outperforms many advanced ensemble methods, according to both the Wilcoxon test and the Friedman test. This method's fundamental flaw is that non-stationary imbalanced data streams cannot be processed [42].

Multimodal data and complicated data with an imbalanced class distribution are frequently used in real-world data analysis. Deep learning is effective in many applications, however, the class imbalance issue has not yet been resolved. Multimodal data learning issues with class imbalance can be handled with NN-CSSVM. For a severely imbalanced data set, it showed a considerable performance gain. The highly nonlinear data are addressed by this cost-sensitive SVM-based prediction function for deep-learning frameworks, which also lessens computing complexity. This deep-learning-based approach with straightforward computations produced improved efficiency compared to standalone SVM models, which need more time for huge data sets. The computational effectiveness of the suggested solution took major significance as the data quantity and dimensionality rose. It has the potential to outperform Cost-Sensitive SVM (CSSVM), Naive Base SVM (NBSVM), and Cost-Sensitive Multilayer Perception (CSMLP) [43].

A considerable improvement in estimating the 5-year survival of breast cancer patients is represented by a hybrid strategy that incorporates SMOTE and PSO. It shows prominent results in terms of improving patient prognosis in the field of oncology because of its capacity to deal with class imbalance, optimize feature selection, and boost prediction accuracy. Particularly for high-dimensional data sets, PSO application for feature selection might increase computational complexity. For effective execution, there must be sufficient computational resources. Performance optimization requires effective parameter tuning for both SMOTE and PSO. The parameters of the algorithm require skill and time. The accuracy of the classifier depends on how good and representative the input data are. The performance of the model could be harmed if the data set has errors or is missing crucial data [44].

Table 6 will provide a comparative analysis of all these hybrid approaches.

Based on the AUC and F1 scores, let us examine each of the aforementioned hybrid approaches in more detail. Taking into account the AUC and F1 score assessed on two distinct data sets Abalone and yeast, taken from UCI [27] data repository and listed in Table 8.

Figure 2 indicates that with an AUC of 0.77 and a fantastic F1 Score of 0.98, WHMBoost exhibits impressive performance on the Abalone data set. This suggests that WHMBoost delivers great precision and recall while excelling at class distinction. Its performance on the Yeast data set is less, with an AUC of 0.92 but a lower F1 Score of 0.36. This mismatch would suggest that WHMBoost is highly specialized for certain data features, like those in the Abalone data set, but may not generalize well to others. With an AUC of 0.84 on Abalone and 0.78 on Yeast, POSENS performs rather well on both data sets. Its F1 Score scores on both data sets, 0.9 and 0.76, respectively, show a balance between precision and recall. This shows that POSENS is a reliable method with consistent performance across many data sets (Fig. 3).

NNCSSVM exhibits stability in its performance, obtaining an AUC of 0.83 on abalone and 0.88 on yeast. A balanced trade-off between precision and recall can be seen by its F1 Score values of 0.8 on both data sets. This indicates that NNCSSVM is a trustworthy option for binary classification jobs with average performance. SMOTE+PSO achieves balanced performance on both data sets with an AUC of 0.8. A decent mix between precision and recall may be seen in its F1 Score scores of 0.82 on abalone and 0.76 on yeast. This shows that SMOTE+PSO is a reliable performer but may not perform well in data sets with extreme imbalance (Table 7).

After going through all the approaches discussed above, it can be concluded that there are many advantages as well as many disadvantages of every strategy, it is not possible to conclude that a particular method works best for handling imbalanced problems. As this is a very complex problem, each strategy can handle only a few of its selected problems. A comparison chart is given in Table 8 which will provide a brief summary of data-level, algorithm-level, and hybrid-level approaches along with their advantages and disadvantages, potential applications, and significant factors.

Table 6 Comparison table for hybrid-level approaches

Approaches	Data sets	Evaluation metrics	Baseline Algorithm	Findings	Limitations
WHMBoost [41]	31 data sets from the KEEL [52] and 9 data sets from HDDT data repository with imbalance rates ranging from 1.87 to 129.44	AUC F-Measure G-Mean	Adaboost RUSBoost RBBoost RHSBoost SMOTEBoost CUSBoost MEBoost	Better performance than single sampling method	Limited to binary class only Not efficient
POSENS [42]	35 data sets from KEEL [52] and UCI [27] repositories	AUC AUPRC F1-M F1-m MC	Bag boost Rusbag Rusboost Smotebag Smoteboost Rbbag Gsmote	It lowers the chance of introducing noises It lowers the chance of introducing sensitivity	Not suitable for non-stationary data sets
NN-CSSVM [43]	Real-world imbalanced data sets from KEEL [52]	AUC G-Mean	CSSVM NBSVM CSMLP	Suitable for data sets having High IR It can address highly nonlinear data Low computational complexity	Not efficient for data sets having less IR
SMOTE + PSO [44]	Breast cancer patients Data of the years 1973–2007 from SEER [55] data repository with 973,125 cases and 118 variables	Accuracy Sensitivity Specificity G-mean	Single LR, C5, 1-nn, and PSO + LR, PSO + C5, PSO + 1-nn without using SMOTE	Improved predictive accuracy Superior robustness	Parameter Tuning Computational complexity Limited data set

Table 7 AUC and F1 score of the hybrid approaches

Approaches	Data set	AUC	F1 score
WHMBoost [41]	Abalone	0.77	0.98
WHMBoost [41]	Yeast	0.92	0.36
POSENS [42]	Abalone	0.84	0.9
POSENS [42]	Yeast	0.78	0.76
NN-CSSVM [43]	Abalone	0.83	0.8
NN-CSSVM [43]	Yeast	0.88	0.87
SMOTE+PSO [44]	Abalone	0.8	0.82
SMOTE+PSO [44]	Yeast	0.87	0.76

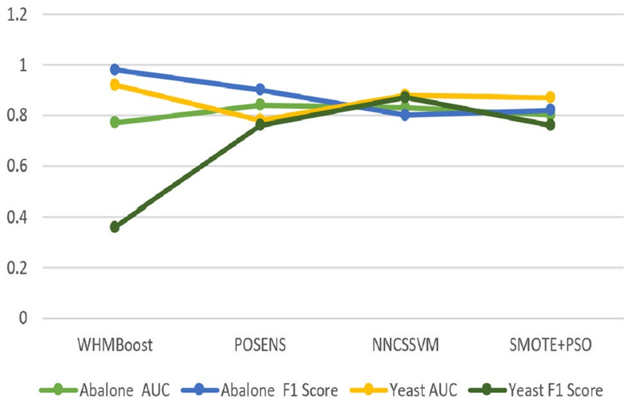

Fig. 3 Performance measured based on AUC and F1 score

The specific characteristics of the data set and the goals of the machine learning task should be taken into consideration when deciding which of these approaches to use. It is crucial to keep in mind that a "Hybrid-Level Strategy" frequently requires more sophistication and parameter adjusting but might produce greater results in severe imbalanced conditions.

Open issues and future work: There are certain things that we have observed and that can still be improved, and they are as follows:

- Compared to multi-class imbalance, the binary class has received more research attention. Because of this, more research may be done to solve the multi-class challenge, and it will also provide a platform for fresh researchers to generate new ideas. In addition, thinking about a strategy that works on both binary and multi-class imbalances may be a good way to proceed.
- Most methods show promising results on one or two data sets, but they do not consistently produce effective

Table 8 Comparison chart on data-level, algorithm-level, and hybrid-level approaches

Aspect	Data level	Algorithm level	Hybrid level
Primary focus	Resampling of data	Modification of algorithm	Combination of data resampling and algorithm modification
Objective	Balance class distribution	Adjust algorithm behavior	Leverage the strengths of both approaches
Strengths	Effective for addressing severe class imbalance Improves recall for the minority class	Precision–recall trade-off can be controlled Fine-tunes the existing algorithms for handling imbalanced data	Comprehensive approach Enhanced model robustness Improved model flexibility
Weakness	Oversampling may introduce noise Undersampling may lead to information loss	Limited to the capabilities of the base algorithm Domain-specific tunning is required	Precise parameter adjustment is necessary Computational difficulty
Key consideration	Selection of the resampling method Potential noise in oversampling Effect on the generalization of the model	Selection of suitable algorithms Adjustment of decision thresholds A trade-off between recall and precision	Model adaption and data preprocessing trade-offs must be balanced Ensuring that different strategies work together
Controlling data set shift	Limited capacity for adaptation	A certain amount of adaptation	Improved adaptability
Overfitting risk	Risky due to data modifications	Lower risk due to algorithm-based adjustments	Moderate risk due to combined approach
Metrics for performance evaluation	Precision Recall F1-score AUC ROC	Precision Recall F1-score AUC ROC	Precision Recall F1-score AUC ROC
Applied fields	Medical diagnostics Fraud detection Anomaly detection	Recommender systems Text classification Image classification	Healthcare analytics Network security Credit scoring
Implementation difficulty	Moderate	Moderate	Moderate to high

results on every data set. Therefore, a strategy that works effectively with all data sets, regardless of IR or size is needed.

- Time efficiency is a crucial consideration. Most procedures show good results but take more time to complete. Time efficiency is a promising area for future research. Therefore, to create a method that is both time efficient and produces positive results, it is essential to keep time efficiency in mind when doing new research.
- Real-world data sets are cluttered with noise. Although some work has been done on the basis of noise so far, there is still room for improvement in this area. Even with noisy data, there can be a method to reduce noise or create excellent results.
- Within the data set, there may be instances where class boundaries overlap. Moreover, it causes issues. Despite the fact that some research has been conducted based on this overlapped area, more work can still be done in this field, because there has not been an effective method developed yet.
- A difficulty with parameters can be found in the algorithm-level approaches. To create a better algorithm, many parameters are considered. The parameters can be manually set for some algorithms whereas they can be pre-configured for others. Since it is difficult to design an algorithm based on all the parameters, this is a field where additional research may be done and the parameter problem can be resolved.

Conclusion

This study has covered a variety of data levels, algorithm levels, and hybrid remedies for class imbalance. After going through all the approaches it has been observed that current research on the high-class imbalance in a massive data environment has severe flaws. After giving a closer look, it is clear that the work in recent years has focused on every category that is data-level, algorithm-level, and hybrid method but Whatever strategy is used, a few facts that all of them have in common like most of them are only useful for some specific set of data. Most of the approaches can not handle high-dimensional data well. Some of the approaches cause class overlapping when data sets are large enough. When it comes to efficiency, certain approaches are prohibitively expensive, which makes them difficult to implement in practice. When it comes to imbalance, data collection might be affected by a number of factors. It is not easy to develop an approach by considering all the factors together, so some factors are ignored in most of the approaches. Keeping these factors in mind, many times such parameters are taken which are too weak, or important parameters may be missed due to which the approach does not work properly or fails to produce a balanced data set. Speed is one of the most essential factors that distinguish one method from others. In this study, it is discovered that certain approaches are extremely slow, taking a long time to provide results and so failing to satisfy expectations.

Author Contributions Study conception, design, and analysis: ZA; Draft manuscript preparation: ZA. Supervised by: SD. All authors reviewed the article and approved the final version of the manuscript.

Funding None.

Availability of Data and Materials Data sharing is not applicable to this article

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Lee Z-J, Lee C-Y, Chou S-T, Ma W-P, Ye F, Chen Z. A hybrid system for imbalanced data mining. Microsyst Technol. 2020;26(9):3043–7.
2. Kamal S, Ripon SH, Dey N, Ashour AS, Santhi V. A mapreduce approach to diminish imbalance parameters for big deoxyribonucleic acid dataset. Comput Methods Programs Biomed. 2016;131:191–206.
3. Arun K, Jabasheela L. Big data: review, classification and analysis survey. Int J Innov Res Inf Secur (IJIRIS). 2014;1(3):17–23.
4. Triguero I, Galar M, Vluymans S, Cornelis C, Bustince H, Herrera F, Saeys Y. Evolutionary undersampling for imbalanced big data classification. In: 2015 IEEE congress on evolutionary computation (CEC). IEEE; 2015. p. 715–22.
5. Ali A, Shamsuddin SM, Ralescu AL. Classification with class imbalance problem. Int J Adv Soft Comput Appl. 2013;5(3):176–204.
6. Kesavaraj G, Sukumaran S. A study on classification techniques in data mining. In: 2013 fourth international conference on computing, communications and networking technologies (ICCCNT). IEEE; 2013. p. 1–7.
7. Koturwar P, Girase S, Mukhopadhyay D. A survey of classification techniques in the area of big data (2015). arXiv:1503.07477.
8. Kaur P, Gosain A. Issues and challenges of class imbalance problem in classification. Int J Inf Technol. 2018;14(1):539–45.
9. Madasamy K, Ramaswami M. Data imbalance and classifiers: impact and solutions from a big data perspective. Int J Comput Intell Res. 2017;13(9):2267–81.
10. Leevy JL, Khoshgoftaar TM, Bauder RA, Seliya N. A survey on addressing high-class imbalance in big data. J Big Data. 2018;5(1):1–30.
11. Hasanin T, Khoshgoftaar TM, Leevy JL, Bauder RA. Severely imbalanced big data challenges: investigating data sampling approaches. J Big Data. 2019;6(1):1–25.
12. Fernández A, Río S, Chawla NV, Herrera F. An insight into imbalanced big data classification: outcomes and challenges. Complex Intell Syst. 2017;3(2):105–20.
13. Rout N, Mishra D, Mallick MK. Handling imbalanced data: a survey. In: International proceedings on advances in soft computing, intelligent systems and applications. Springer; 2018. p. 431–43.
14. Lemnaru C, Potolea R. Imbalanced classification problems: systematic study, issues and best practices. In: International conference on enterprise information systems. Springer; 2011. p. 35–50.
15. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016;5(4):221–32.
16. Ahmed Z, Askari SMS, Das S. Comparative analysis of recent data-level methods for imbalance classification. In: 2023 4th international conference on computing and communication systems (I3CS). IEEE; 2023. p. 1–6.
17. A gentle introduction to imbalanced classification. https://machinelearningmastery.com/what-is-imbalanced-classification/. Accessed 26 Oct 2021.
18. Abdi L, Hashemi S. To combat multi-class imbalanced problems by means of over-sampling techniques. IEEE Trans Knowl Data Eng. 2015;28(1):238–51.
19. Somasundaram A, Reddy US. Data imbalance: effects and solutions for classification of large and highly imbalanced data. In: International conference on research in engineering, computers and technology (ICRECT 2016). 2016. p. 1–16.
20. He H, Ma Y. Imbalanced learning: foundations, algorithms, and applications. 2013.
21. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
22. Bej S, Davtyan N, Wolfien M, Nassar M, Wolkenhauer O. Loras: an oversampling approach for imbalanced datasets. Mach Learn. 2021;110(2):279–301.
23. Kowalski BR, Bender C. k-nearest neighbor classification rule (pattern recognition) applied to nuclear magnetic resonance spectral interpretation. Anal Chem. 1972;44(8):1405–11.
24. Kramer O, Kramer O. Scikit-learn. Machine learning for evolution strategies. 2016. p. 45–53 .
25. Vuttipittayamongkol P, Elyan E. Neighbourhood-based undersampling approach for handling imbalanced and overlapped data. Inf Sci. 2020;509:47–70.
26. KEEL: a software tool to assess evolutionary algorithms for Data Mining problems (regression, classification, clustering, pattern mining and so on). https://sci2s.ugr.es/keel/datasets.php. Accessed 03 July 2022.
27. UCI Machine Learning Repository: Data Sets. https://archive.ics.uci.edu/ml/datasets.php?format= &task=cla &att= &area= &numAtt= &numIns= &type= &sort=nameUp &view=list. Accessed 03 Sept 2022.
28. Vo MT, Nguyen T, Vo HA, Le T. Noise-adaptive synthetic oversampling technique. Appl Intell. 2021;51(11):7827–36.
29. Arefeen MA, Nimi ST, Rahman MS. Neural network-based under sampling techniques. IEEE Trans Syst Man Cybern Syst. 2020;52(2):1111–20.
30. Liu C, Jin S, Wang D, Luo Z, Yu J, Zhou B, Yang C. Constrained oversampling: an oversampling approach to reduce noise generation in imbalanced datasets with class overlapping. IEEE Access. 2020;10:91452–65.
31. Krawczyk B, Bellinger C, Corizzo R, Japkowicz N. Undersampling with support vectors for multi-class imbalanced data classification. In: 2021 international joint conference on neural networks (IJCNN). IEEE; 2021. p. 1–7.

32. Sáez JA, Luengo J, Stefanowski J, Herrera F. Smote-ipf: addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. Inf Sci. 2015;291:184–203.
33. Vuttipittayamongkol P, Elyan E, Petrovski A, Jayne C. Overlap-based under sampling for improving imbalanced data classification. In: International conference on intelligent data engineering and automated learning. Springer; 2018. p. 689–97.
34. Liu J. Fuzzy support vector machine for imbalanced data with borderline noise. Fuzzy Sets Syst. 2021;413:64–73.
35. Yuan B-W, Luo X-G, Zhang Z-L, Yu Y, Huo H-W, Johannes T, Zou X-D. A novel density-based adaptive k nearest neighbor method for dealing with overlapping problem in imbalanced datasets. Neural Comput Appl. 2021;33(9):4457–81.
36. Tanveer M, Sharma A, Suganthan PN. Least squares knn-based weighted multiclass twin svm. Neurocomputing. 2021;459:454–64.
37. Kumbure MM, Luukka P, Collan M. A new fuzzy k-nearest neighbor classifier based on the Bonferroni mean. Pattern Recognit Lett. 2020;140:172–8.
38. Lin E, Chen Q, Qi X. Deep reinforcement learning for imbalanced classification. Appl Intell. 2020;50(8):2488–502.
39. Tao X, Li Q, Ren C, Guo W, He Q, Liu R, Zou J. Affinity and class probability-based fuzzy support vector machine for imbalanced data sets. Neural Netw. 2020;122:289–307.
40. Boosting methods for multi-class imbalanced data classification: an experimental review. https://journalofbigdata.springeropen.com/track/pdf/10.1186/s40537-020-00349-y.pdf. Accessed 03 May 2022.
41. Zhao J, Jin J, Chen S, Zhang R, Yu B, Liu Q. A weighted hybrid ensemble method for classifying imbalanced data. Knowl Based Syst. 2020;203: 106087.
42. Zhang J, Wang T, Ng WW, Pedrycz W. Ensembling perturbation-based oversamplers for imbalanced datasets. Neurocomputing. 2022;479:1.
43. Kim KH, Sohn SY. Hybrid neural network with cost-sensitive support vector machine for class-imbalanced multimodal data. Neural Netw. 2020;130:176–84.
44. Wang K-J, Makond B, Chen K-H, Wang K-M. A hybrid classifier combining smote with pso to estimate 5-year survivability of breast cancer patients. Appl Soft Comput. 2014;20:15–24.
45. Huang J. Performance measures of machine learning. University of Western Ontario. 2008.
46. He H, Bai Y, Garcia EA, Li S. Adasyn: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE World Congress on Computational Intelligence). IEEE; 2008. p. 1322–28.
47. Farquad MAH, Bose I. Preprocessing unbalanced data using support vector machine. Decis Support Syst. 2012;53(1):226–33.
48. Han H, Wang W-Y, Mao B-H. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: Advances in intelligent computing: international conference on intelligent computing, ICIC 2005, Hefei, China, August 23–26, 2005, Proceedings, Part I 1. Springer; 2005. p. 878–87.
49. Nguyen HM, Cooper EW, Kamei K. Borderline over-sampling for imbalanced data classification. Int J Knowl Eng Soft Data Paradig. 2011;3(1):4–21.
50. scikit-learn: machine learning in Python—scikit-learn 1.3.0 documentation. https://scikit-learn.org/stable/. Accessed 17 Sept 2023.
51. SIGKDD: KDD Cup 2008: Breast cancer. https://kdd.org/kdd-cup/view/kdd-cup-2008. Accessed 17 Sept 2023.
52. KEEL: a software tool to assess evolutionary algorithms for Data Mining problems (regression, classification, clustering, pattern mining and so on). https://sci2s.ugr.es/keel/datasets.php. Accessed 17 Sept 2023
53. LIBSVM data: classification, regression, and multi-label. https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/. Accessed 17 Sept 2023.
54. Find Open Datasets and Machine Learning Projects | Kaggle. https://www.kaggle.com/datasets. Accessed 17 Sept 2023.
55. SEER Incidence Data, 1975–2020. https://seer.cancer.gov/data/. Accessed 19 Sept 2023.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.