Multimedia Tools and Applications (2024) 83:63243–63290
https://doi.org/10.1007/s11042-023-17864-8

Class overlap handling methods in imbalanced domain: A comprehensive survey

Anil Kumar[1] · Dinesh Singh[1] · Rama Shankar Yadav[1]

Received: 27 September 2022 / Revised: 18 September 2023 / Accepted: 13 December 2023 /
Published online: 11 January 2024

Abstract

Class overlap in imbalanced datasets is the most common challenging situation for researchers in the fields of deep learning (DL) machine learning (ML), and big data (BD) based applications. Class overlap and imbalance data intrinsic characteristics negatively affect the performance of classification models. The data level, algorithm level, ensemble, and hybrid methods are the most commonly used solutions to reduce the biasing of the standard classification model towards the majority class. The data level methods change the distribution of class instances thus, increasing the information loss and overfitting. The algorithm-level methods attempt to modify its structure which gives more weight to the misclassified minority class instances in the learning phases. However, the changes in the algorithm are less compatible for the users. To overcome the issues in these methods, an in-depth discussion on the state-of-the-art methods is required and thus, presented here. In this survey, we presented a detailed discussion of the existing methods to handle class overlap in imbalanced datasets with their advantages, disadvantages, limitations, and key performance metrics in which the method shown outperformed. The detailed comparative analysis mainly of recent years' papers discussed and summarized the research gaps and future directions for the researchers in ML, DL, and BD-based applications.

Keywords Class overlap · Class imbalance · Deep learning · Machine learning · Big data · Re-sampling

1 Introduction

Now a days class overlap in imbalanced datasets is a common and challenging situation for machine learners. The datasets in real-world applications are suffering from class imbalance

✉ Anil Kumar
anilk@mnnit.ac.in

Dinesh Singh
dinesh_singh@mnnit.ac.in

Rama Shankar Yadav
rsy@mnnit.ac.in

[1] Department of Computer Science and Engineering, Motilal Nehru National Institute of Technology Allahabad, Prayagraj, UP, India

and overlap due to limitations in data collection. In recent years researchers found that class overlap is a painful and more problematic situation even if a dataset is balanced. It degrades the performance of the standard classification models context to machine learning (ML) [1–3] and deep learning (DL) [4, 5] in small, medium, and big data (BD) [6–12] domains. Class overlapping is a region where multiple class instances are shared in a common region in data space. The reason behind class overlapping is that data instances have similar feature values but belong to different classes. The performance of the standard learning algorithm decreases in both situations i.e. imbalanced and overlapped data. The joint impact of both imbalance and overlap on the performance is more degrading in comparison to individuals [2]. The existing learning algorithms are not able to train effectively in overlapped regions because of poor visibility of minority instances. In this situation, majority class instances are more dominant in the overlapped regions and frequently visible to the learner in comparison to minority instances. This problem tends to transfer choice boundary toward the majority class and thus, high misclassification of the minority class instances. It increases the false positive rate which is a harsh situation and undesirable in real problems especially in the field of medical science [13–18], anomaly detection [19–22], machinery fault detection [23–33], business management [34–38], financial sector [39–42], fake spam analysis [43–45], Data mining and text classification [46–54], computer vision and image processing [55–62], bio-informatics [63–69], Detection of faulty software modules [70–72], and information & cybernetic security [73–78] etc.

The existing solutions discussed in the literature for handling overlapped and imbalanced datasets were categorized based on class distribution or overlap of class instances [2, 38, 79–83]. The instance distribution-based method resolved the issue by resampling class instances i.e. by performing under, over, or hybrid sampling of instances. These methods suffer from loss of informative instances in case of under-sampling and increase overfitting in oversampling. Class overlapping-based methods work in two phases. Identification of overlapping region is performed in the first phase and instances in the overlapped region are handled in the second phase. There are three approaches, discarding overlapped regions, merging overlapped regions, and separating overlapped regions [84]. In the discarding-based approach overlapped region is ignored and the algorithm learns in a non-overlapped region. In the merging approach overlapped regions are considered as a new class level and handled using two-tier architecture classification models. In the upper tier of the structure, the model focuses on the entire dataset with one additional new class label ("overlap_class") of the overlapped region. If the test sample belongs to the overlapped region which is identified by the upper-level model, then the model used in the lower tier predicts the actual class of the sample. In separating overlapped approaches two different models are needed for learning and testing purposes. One model is used for separating entire data into the overlapped and non-overlapped regions and the second model is used for learning both regions separately. The K-Nearest Neighbor (KNN) is used for separating data into overlap and non-overlap regions and two support vector machines (SVMs) are used for learning each region separately (Fig. 1).

Imbalance and overlap are key problems in machine learning [3, 79, 80, 82], deep learning [85–87], and big data [7, 88, 89]. Today DL and BD become the fastest growing research areas where new architectures are added very frankly and CNN become the most popular classification model for time series prediction and analysis [90]. The year-wise number of publications included in this survey is shown in Fig. 2. This survey presents a detailed overview of the majority of recent year methods based on DL, and ML context in small, medium, and BD domains. The summarized structure of the survey is shown in Fig. 1.

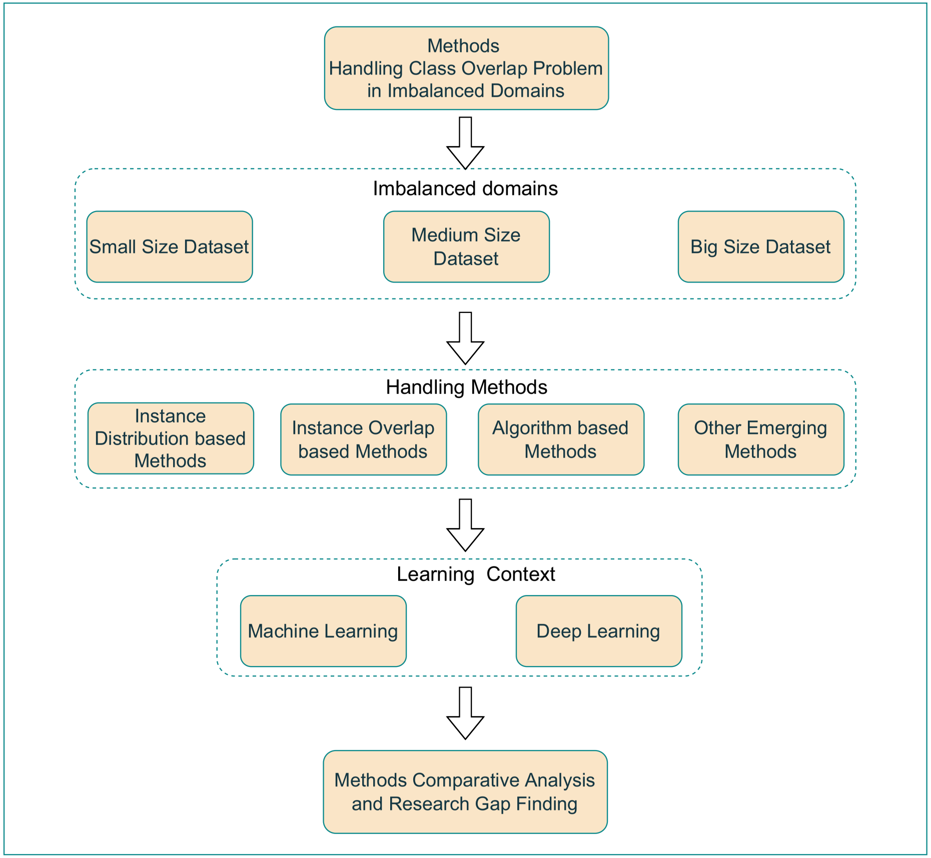

Fig. 1 Structure of our survey based on covered overlap in imbalance domain, learning method, and environment

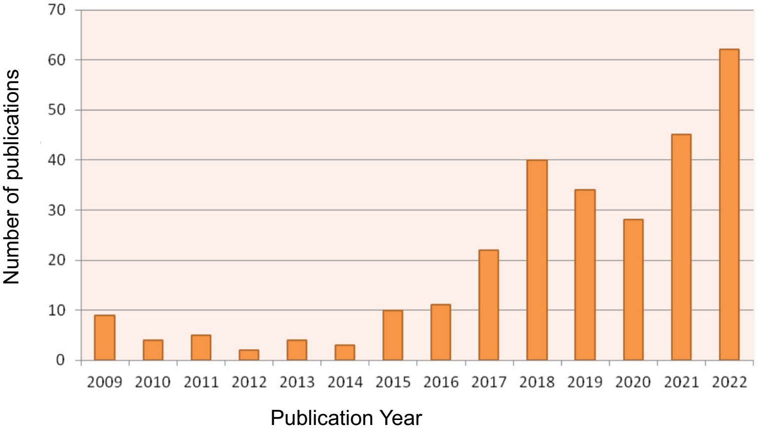

Fig. 2 Number of publications based on class overlap and imbalance in machine learning, deep learning, and big data context

1.1 Motivation

Class imbalanced and overlapped datasets are bound to bias the performance of the standard machine learning algorithms and pull off good results towards the majority. But if any domain in which minorities are a more interesting pattern as compared to the majority then it will be a crucial problem because in such applications misclassification of minority class instance is dangerous for example in medical diagnostics if the prediction of COVID-19 positive (minority) patients as negative (majority). In many such applications like medical diagnostics, fraud detection, defect prediction, etc., approximately a hundred percent minority accuracy is expected. Many recent studies showed that the presence of overlapping is more dangerous even if data is balanced. Many methods and frameworks have been proposed by researchers in recent years. Occurrences of class overlapping and imbalance are natural at the time of data generation and preparation and cannot be avoided. Therefore an efficient method is required to handle class overlaps. A huge number of solutions for handling classes overlapping in the imbalanced domain are continuously being proposed by researchers. It is quite difficult to select a feasible solution and method for such real-world applications. The proposed methods may not be suitable to tackle problems in all types of datasets globally, therefore it is still required and in demand to develop an optimal solution that may have the ability to stop the biasing nature of algorithm performances towards the majority class.

1.2 Why this survey?

The existing surveys [2, 38, 79–84] only explored class overlapping issues in specific domains i.e. machine learning, deep learning, and big data problems in limited applications. Here in the present work we briefly summarized issues, research gaps, and future directions in extreme imbalance domain-based applications where class overlapping problems have been taken into consideration. Major challenges in class overlapping domains are under consideration at the data analysis level, preprocessing, algorithm level, and meta-learning environment. For the data analysis purpose important and core focusing terms are multi-classification, binary classification, and singular problems. Open challenges at the data preprocessing level are undersampling, oversampling, hybrid sampling, feature extraction, and selection. At the algorithm level, crucial challenges are specialized strategies, hyperparameter, and cost threshold. Meta-learning barriers are recommended ensemble methods, classifiers, and data resampling strategies.

Our major contributions in this paper are outlined as follows:

1. Presented most used and recent year methods for handling class overlapping and imbalanced problems context to machine learning and deep learning in small, medium, and big data domains.
2. Briefly discussed category-wise methods handling class overlap and imbalance with advantages, disadvantages, limitations, and issues.
3. Presented comparative analysis of past year surveys based on important parameters such as imbalance, overlap, advantages, limitations, research gap, and performance metrics.
4. Outlined summary of detailed research gaps, issues, and challenges in overlap and imbalance handling methods for researchers in context to machine learning, and deep learning in small, medium, and big data problems.
5. Apart from these above the survey also presents recent imbalanced and overlapped application domains along with their major related research papers for the current researchers in these domains.

The remaining parts of the survey are organized in the following order: Section 2, presented data intrinsic difficulties such as overlap and imbalance, overlap- -characterization task, application domains, and important performance evaluation metrics of the classifiers used to analyze class overlapping effects. In Section 3, we addressed the last five years' literature methods and categorized them based on techniques used in ML, DL, and BD environments. In Section 4, we presented a comparative analysis of the existing surveys. In Section 5, we presented a performance metric summary of the existing methods and summarized research gaps. At last in Section 6, we concluded and discussed the future direction of the researcher.

2 Dataset intrinsic difficulties, applications, and performance evaluation

This section presents the class overlap and imbalanced domains where these datasets' intrinsic properties need to be tackled and draw attention to the data science and machine learning domains researchers. In these application domains, the class of interest patterns belongs to the minority class instances and has a high penalty for misclassification errors. In the remaining part of this section, the imbalanced application domain along with recent related articles and important metrics for measuring the performance of classification models in the unbalanced learning domain have been discussed.

2.1 Dataset intrinsic difficulties

Generally compiled data with unequal class distribution degrades the performances of standard learning models [91]. On the other hand, it is well known that class imbalancement is not only responsible for this unexpected behavioral nature of real-world dataset [92]. Many other factors are solely responsible for degrading the performance of the learning models. These factors are the data intrinsic characteristics (class overlap and imbalance) [93, 94], data irregularities (missing-values, small-disjuncts) [93], and data difficulties (data-lacking, data-noise, data- shifting) [95]. The most harmful problem that has been characterized by data science researchers is class overlapping [2, 79, 80]. The combined effects of both imbalance and overlap have been an extremely hot topic for researchers for decades in several domains such as medical, commuter vision & image processing, security, etc. [2, 79, 80, 96, 97].

2.1.1 Class imbalance

The datasets in which class instances are not equally distributed is called imbalanced dataset The class imbalance in the dataset creates more problems when minority instances are very less and have comparatively high misclassification costs. The majority and minority classes are known as negative and positive classes respectively. The imbalance is measured by imbalance ratio (IR), which is defined as the ratio of the counts of the majority class instances and the counts of minority class instances as given in (1). The minority percent is computed by (2) [2, 3]. Class imbalance is illustrated in Fig. 3, where the blue star represents the majority class instance and the red triangle represents minority class instances.

$$Imbalance\ Ratio(IR) = \frac{Counts\ Majority\ Class\ Instances}{Count\ Minority\ Class\ Instances} \tag{1}$$

$$Minority(\%) = \frac{Counts\ Minority\ Class\ Instances}{Count\ Majority\ Class\ Instances} * 100 \tag{2}$$

2.1.2 Class overlapping

The dataset in which instances of multiple class shares a common region in the data space is called an overlapped dataset as illustrated in Fig. 3. These instances are similar in some or all feature values but belong to a different class, and such a problem is a considerable barrier in classification tasks during training and learning. In the overlap region, the majority class becomes dominant because of the highly visible to the learners in comparison to the minority. This causes the choice margin for prediction to transfer toward the majority thus, leading to the miss-classification of minority instances closer to the margin of overlap, which is not expected in real applications. The amount of class overlap area was not well-formulated [98], and a global dimension to measure overlap percentage is undefined. Many methods have been proposed to estimate the overlapped percentages but with restrictions. In [99], the overlap percentage is estimated from the ratio of overlap area to entire data space. Another popular approach is using the classification error use to estimate the overlap percentage miss-classified by the KNN [100–102]. However other techniques were proposed in the past year by researchers but they cannot be used globally [98, 103, 104]. Another method proposed in [105] is based on support vector data description (SVDD). The SVDD shrinks the spherical class boundary towards its centroid subject to minimize the radius and then counts the number of instances that are common in both classes. This is applicable only for spherical data shapes. The class overlapping percentage is estimated by the formula given in (3) [2].

$$Overlapping(\%) = \frac{Number\ of\ instances\ in\ overlapped\ region}{Number\ of\ instances\ in\ minority\ region} * 100 \tag{3}$$

2.1.3 Class overlap characterizing components

Identifying and characterizing class overlap in an imbalanced learning domain is still a trouble and tight spot for researchers, since there is no standard and clear well-defined quantification

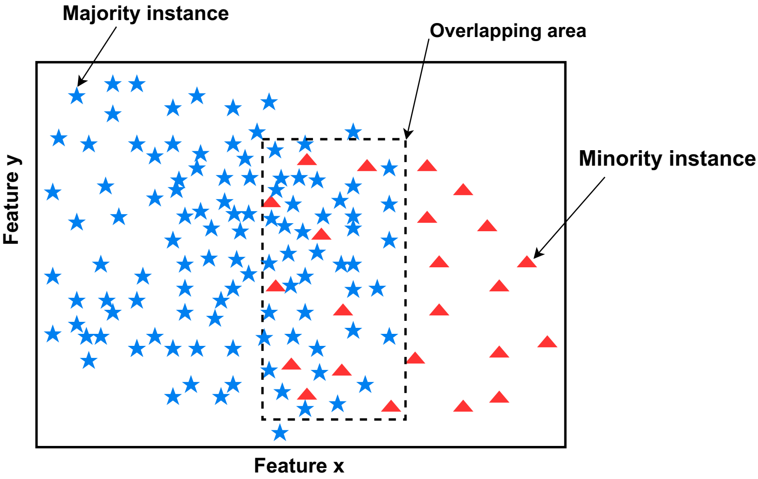

Fig. 3 Two-dimensional space representation of class overlap in imbalanced datasets by two random features x and y. The Blue star denotes the majority class instance and the red triangle denotes the minority class instance. The rectangular shaded area represents an overlapping region

and measurement of this hot topic for real-world application domains [1, 2, 79]. The class overlapping problem can be characterized using three primary consecutive tasks i.e. data domain decomposition, overlap identification, and quantifying class overlap. The decomposition task crumbles entire data into regions of attention for learning and prediction. The region identification task spots the overlap regions. The quantification task measures the amount of overlapped regions. Class overlap may be characterized in different perspectives based on approaches applied to these primary tasks thus leading to different representations of the overlap problem. Thus individual representations measure overlap differently according to the application domains. All the existing approaches for measuring class overlap have three major essential components as shown in Fig. 4.

- **Decomposition of data domain into interesting regions:** The Three most used approaches based on the distance for dividing entire data feature space into interesting regions are statistical, geometrical, and graph-based distances. Statistical distance approach based on the distance between class distributions for example Fisher Linear Discriminant. The geometrical distance approach is based on the distance between pairs of class instances for example Euclidean distance. Graph-based distance approach based on the geodesic distance for example Minimum Spanning Tree.
- **Identification of interesting regions:** The major approaches for identifying overlap regions are: Discriminative analysis is used for determining the distinctive capability of the features, and in-depth class distribution characteristics need to be analyzed. Interesting regions are observed in the area where classes are overlapped and maximize class separability. Feature space division is used for dividing feature space into discrete intervals for analyzing the characteristics of data. The problematical regions are encircled in the specific range of feature space. Neighborhood search is based on k-nearest neighbor (KNN) search and interested regions are those that have maximum error generated by the KNN classifier. Hypersphere coverage the informative and necessary class instances are covered by subsets and problematic regions are on all sides of the hypersphere. The min-

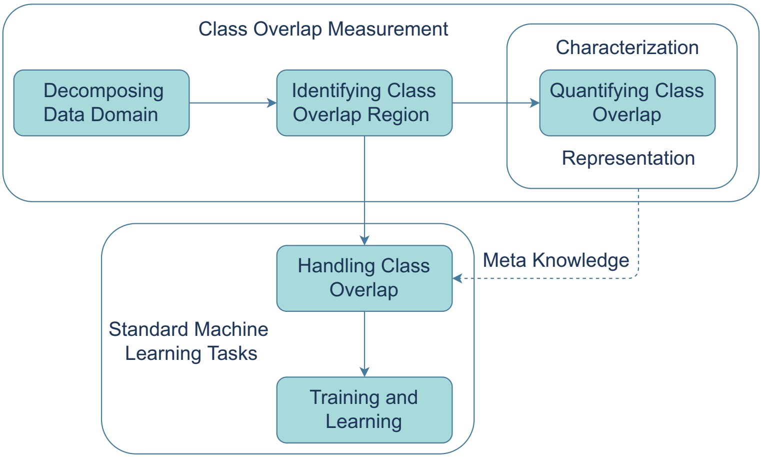
Class Overlap Measurement
Characterization
Decomposing Data Domain
Identifying Class Overlap Region
Quantifying Class Overlap
Representation
Meta Knowledge
Handling Class Overlap
Standard Machine Learning Tasks
Training and Learning

Fig. 4 Block diagram of the class overlap characterization measurement in imbalanced dataset in machine learning, deep learning, and big data domain

imum spanning tree is used to represent the entire data spanning tree and problematical regions are those connected vertices that do not satisfy membership degrees.

- **Quantifying overlap area:** Finally this component is used to measure the quantity of the interesting regions. The four major interested region representations are based on feature, instance, structural, and multi-resolution overlaps.

2.2 Applications

Class overlapping and imbalance domains have drawn attention from Researchers over the years because modern datasets are inherently prone to these problems. Figure 5 shows application areas based on our literature search. The detailed domain area along with recent related research references are given below:

- **Medical science:** Machine learning tools have proved to be very effective in the diagnosis of some life-threatening diseases. However, the number of patients suffering from the disease in the data set used to train the machine-learning model is much less than normal people. And there is more similarity in the symptoms of patients and normal people such as cancer, heart failure, etc [13–18].
- **Machine fault detection:** Rolling bearing is one of the most important components in a rotating machine. Machine learning-based fault detection is the most commonly used method which plays a vital role in reliable and safe operation. The most common issue is severe data imbalance because the frequency of a faulty component is much lower than that of a non-faulty one [23–33].
- **Stock market:** Current research and studies show that imbalance and overlap negatively affect stock returns. The experimental result has proven that a one percent trading imbalance, a difference of around seven percent can result in an overall trading return [39–42].
- **Intrusion detection system:** These systems are used to detect suspicious behavioral activities and generate alert messages as an alarm when such incidents happen. The recent research carried out in network intrusion detection systems are: [19–22].

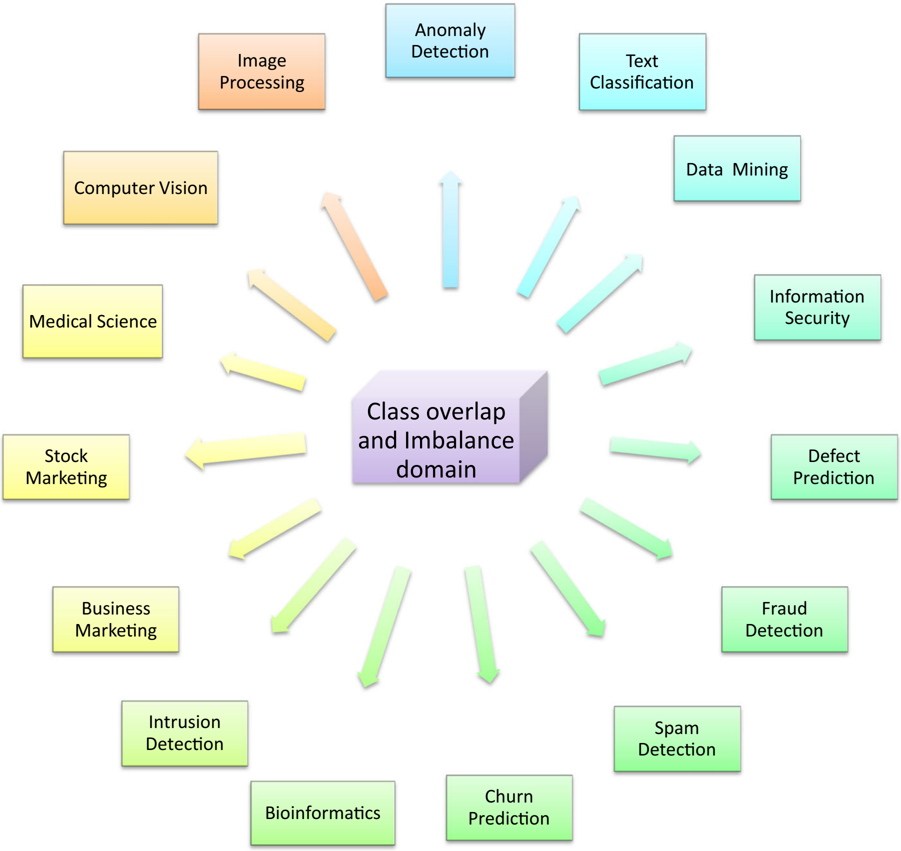

Fig. 5 Imbalance application domains where class overlapping methods have been used to resolve the performance degradation issues of the standard machine learning and deep learning models

- **Fraud detection:** Detecting fraud transactions is a big challenge for credit and debit card companies as there is a high level of behavioral similarities between fraud and normal transactions. Apart from this, occurrences of fraud transactions are much less than normal transactions. Therefore the performance of the machine learning algorithm is biased toward normal transactions due to such intrinsic nature of transactional dataset [8, 106].
- **Software fault detection:** In software development only a small part of modules are faulty and most of them are non-faulty so such data is naturally imbalanced. Most classification models face the problem of working with data with imbalanced class distributions since most such models usually assume that each class has the same misclassification cost [70–72].
- **Computer vision:** Systems are used to retrieve important information about the image including image features, structure, and labels of image data. The imbalance occurs when there is a lack of interesting images in a huge amount of image data [61, 62].
- **Image processing:** The method operates on image data to improve quality and extract useful information for various purposes like medical diagnosis, remote and ultraviolet sensing, robot vision etc.[55–60]

- **Information security:** In the last few years, the use of machine learning algorithms has played an important role in security and authentication purposes. The data used for this purpose is highly imbalanced and overlapped [73–78].
- **Churn prediction:** Customer churning is a rare event in many companies such as telecommunications, credit cards, and many other service providers. In other words, we can say that churners counting is very low as compared to non-churners. A small percentage of customer retention may generate a larger percentage of revenue in such business companies [34–37].
- **Spam detection:** In this domain, the instances of spam reviewers are in the minority as compared to non-spammers. The ML-based filters face great challenges due to this imbalance of distribution while filtering spammers, because of the biasing nature of learning algorithms [43–45].
- **Data mining:** Many applications such as text summarization, novel pattern mining from historical data, time series analysis, web mining, etc. are using machine learning algorithms for prediction. The dataset used for training purposes is imbalanced and overlapped. Since interesting mining patterns are minority class instances, therefore improved and adaptive methods need to tackle these issues [50–54].
- **Text classification:** Class imbalance and larger features are two major problems of textual data classification. there are many areas belonging to this domain such as sentiment analysis, healthcare fraud detection, etc. The feature selection-based techniques have been used by the researcher for handling imbalanced big data in this domain [46–49].
- **Bio-informatics:** It is an essential application tool used for computational analysis to interpret biological information such as Gene therapy, protein structure prediction, new drug discovery, evolutionary and microbial, etc. The machine learning-based application is used to handle data intrinsic characteristics that are imbalance and overlap [63–69].
- **Business marketing:** It allows individual organizations including commercial and Government institutions to sell their products and other services. Machine learning and business intelligence based tools are used to search for customers, increase productivity, and prediction of future demands [107–113].

2.3 Performance metrics

Some performance metrics are not affected by imbalanced instance distributions but many of them are misleading. The most common metrics for the classification of overlapped and imbalanced datasets are sensitivity, specificity, and balanced accuracy(BA), G-mean, Area Under Curve (AUC), and F-score [1–3]. In imbalanced applications, accurate detection of minority class instances is very difficult. This is usually assessed in terms of recall or true positive rate (TPR) and evaluated by using (6), where TP, TN, FP, and FN are taken from the confusion matrix as shown in Table 1. As sensitivity only reflects the performance over one class also relates to another metric, such as specificity i.e. true negative rate as given in (8), where TN and FP represent true negative and false positive respectively.

Table 1 Confusion matrix

Actual Value	Predicted Value	
	Positive Class	Negative Class
Positive Class	True Positive(TP)	False Negative(FN)
Negative Class	False Positive(FP)	True Negative(TN)

Balanced accuracy is defined as the average accuracy of both classes and is also known as balanced mean accuracy, average accuracy, or macro accuracy, etc., and computed by using (9). **Accuracy** computed as given in (4) can mislead in case of highly imbalanced data and the majority class accuracy (TN and TN + FP) are highly dominant. For example, a correctly predicted majority class of 10000 instances with a total misclassified positive class with 100 instances gives an accuracy of 99 percent, which is bad for a good classifier. And similarly, 50% BA reflects the good performance of the model. Therefore it frequently replaces the accuracy metric and becomes the most common performance measure used in an imbalanced case.

G-mean metric is used to assess the overall performance which geometric mean of sensitivity and specificity and evaluated as given in (10). G-mean is most frequently used as an assessing performance metric. Since both are evaluating the average of common parameters sensitivity and specificity both can be used interchangeably. The inequality relation between geometric mean and square root mean is given by formula as in (11), hence similar inequality relationship must be satisfied between balanced accuracy G-mean performance metrics as formulated in (12).

Precision describes the ability of a classification model to identify a relevant instance in testing data instances i.e. a higher precision denotes a higher probability of correct prediction of a positive instance. It is the ratio of positive predicted value and relevant retrieved instances as given in (5). It is also known as a positive predicted value. It answers that among all positive predictions how many are real positives and how many are real negatives but the model incorrectly predicts them as positives. This metric is very useful when datasets are imbalanced. The formula for computing precision is the ratio of true positives to the sum of true positives and false positives.

Receiver Operating Characteristics Curve (ROC) is a plot between TPR on the y-axis and FPR on the x-axis. The curve area between (0, 0) and (1, 1) is called the area under the curve (AUC). A higher AUC value of a model has a better capability to predict an actual positive class as a positive class and an actual negative class as a negative class in comparison to a model having a lower AUC. The minimum threshold value is 0.5 which means the model predicts a positive class to be negative and a negative class to be positive. Two or more models' performances can be compared based on AUC value.

Sensitivity describes the ability of a classification model to identify all data instances in a relevant class i.e. higher sensitivity means the model can predict maximum positive instances as positive. It is defined as the ratio of true positive and the sum of true positive and false negative as given in (6). It describes how efficient the classifier is when predicting a positive class on desired positive outcomes.

Specificity is the true negative rate of the model. It is the ratio of TN and the sum of TN and FP as given in (8). Balance accuracy is the average of sensitivity and specificity and is computed by using (9). The accuracy of the model is the ratio of truly predicted values of target instances and their actual label value in the testing part of the dataset. In many cases generally for imbalance data accuracy may be descriptive i.e. model performance is good for the majority class and poor for the minority class and it is not sufficient to measure the performance of the model. The accuracy is the ratio of correct prediction to the sum of correct and incorrect predictions which are computed by (4):

F- score is used to relate precision and recall together where recall is prioritized over precision, in this case, the F-score of a model can be calculated by using (7). For β =1,F_1 score is the harmonic mean of precision and sensitivity.

Matthews correlation coefficient (MCC) is a more consistent and powerful statistical measure of the classification models used in the imbalance learning domain. The MCC value range between -1 t0 +1, where -1 indicates a total disagreement and +1 represents the total agreement between predicted and actual instance labels. The MCC is computed by using (13).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$F_{\beta} = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{(\beta^2 \cdot Precison + Recall)} \tag{7}$$

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

$$BalancedAccuracy = \frac{Sensitivity + Specificity}{2} \tag{9}$$

$$G - mean = \sqrt{Specificity * Sensitivity} \tag{10}$$

$$inequality : \frac{x + y}{b} \geq \sqrt{xy} \tag{11}$$

$$BalancedAccuracy \geq G - mean \tag{12}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{13}$$

3 Literature review

A detailed discussion about the past year's research in the context of deep, machine learning, and big data is presented in this part of the survey. The reviews are based on the category of methods handling overlap plus imbalances in both DL, ML, and BD. The current survey includes the majority of the past five years of research and major surveys in DL, ML, and BD domains. The first part comprises a detailed discussion of ML and category-wise related work. The first part describes the DL background and related research. The third part presents reviews of big data environments

3.1 Machine learning based review

In previous literature surveys and various reviews, methods of handling imbalanced and overlapped data were grouped based on their convenience and used techniques. Some studies categorized methods as per compatibility whether it was able to handle skewness of data or similarities in feature value. Some studies categorized it based on approaches without any changes in the original data. Some studies cauterized it based on desirable requirements of performance metrics. Many studies concluded that the negative effect of the existing standard classification model is very high in overlapping scenarios in comparison to imbalance. The overlap impacts the performance even if data is balanced [2, 3]. The existing methods deal either with features of datasets or with the instance. Here in our study, we categorized more optimistic ways of the existing methods of handling class overlapping in an imbalanced dataset. The schematic representation of our categorization is shown in Fig. 6.

3.1.1 Data level methods

This includes methods that handle the class overlapping issues by changing the original data distribution and further subdividing it into instance distribution and overlap-based subcategories.

Instance Distribution Based Method These methods are based on data re-sampling (undersampling, oversampling, and hybrid-sampling) of instances of the entire dataset.

The major issues with these methods are excessive elimination of majority instances causes information loss and regenerating synthetic instances creates overfitting. An in-depth summary of the instance distribution based methods is shown in Table 2 along with their pros and cons.

Under-sampling In the undersampling based method majority of instances were eliminated from the entire dataset. Vuttipittayamongkol et al. [114] proposed an under-sampling method based on recursive neighbor searching (URNS). The URNS identifies class overlap regions by searching the KNN of the majority class instances. URNS removes negative instances after pre-processing data by z-score normalization. The KNN searches nearer the surrounding minority instances and then removes majority instances which highly weakens the visibility of minority instances. To prevent excessive elimination URNS is used to target only those majority instances for elimination which weakens the visibility of more than one minority instance. URNS also ensures sufficient elimination by a twice repeated recursive search. URNS ensures high sensitivity. The major benefits of URNS are that under-sampling is independent of imbalance and methods dealing with a class overlap which is the cause of misclassification. Rui et al. [115] proposed an under-sampling method based on the random forest cleaning rule (RFCL). The RFCL eliminated those majority instances that cross newly updated classification boundaries within the margin threshold. The threshold value is estimated by maximizing the f-1 score of the classifier. The method can minimize both overlap and imbalance. The method used a random forest tree for computing the margin value of class instances by (14) where $\mathrm{V}_{\mathrm{true}}$ and $\mathrm{V}_{\mathrm{false}}$ are the numbers of true and false voting by random forest tree respectively. Soumya Goyal proposed a neighborhood-based under-sampling approach for software defect prediction using an artificial neural network. The method reduces both overlapping and imbalancement.

$$margin(X_i) = \frac{V_{true} - V_{false}}{V_{true} + V_{false}} \tag{14}$$

The method calculates the margin value of each instance and then draws a box plot for the majority and minority classes respectively. The overlapping degree is OD= Q_3 (minority) - Q_1 (majority) where Q_3 third quartile of the minority margin box plot and Q_1 is the first quartile of the majority margin box plot. The smaller value of OD represents a higher separation between majority and minority. The methods search margin threshold if the majority instance crosses the margin threshold boundary then it will be eliminated.

Devi et al. [116] proposed a model based on one class SVM and undersampling to handle overlapping effects. The model is designed in two stages i.e. pre-processing and training. Stage one involves the identification of overlapped regions, removing overlapped instances, and under-sampling. The SVM is used to detect outliers and nest stage removing overlapped instances using the Tomek link pair. Finally, three classification models feed-forward neural

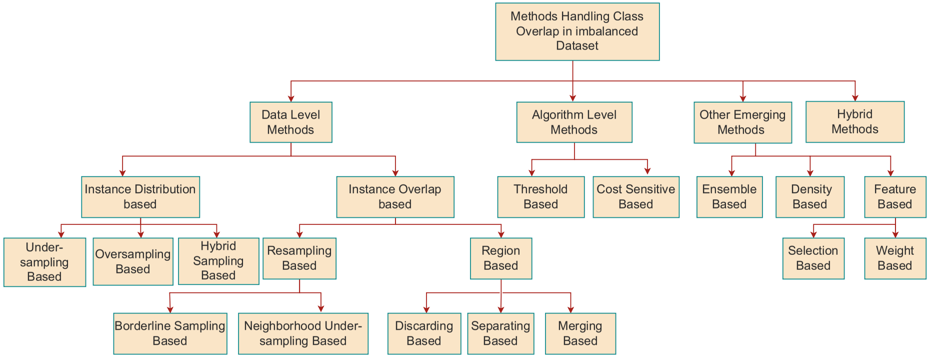

Fig. 6 Categorization of existing methods to handle class overlap in an imbalanced domain. The key characteristics behind classifying overlapping methods are based on approaches and strategies used to handle issues in the recent real-world application domains

Table 2 Overview of instance distribution based methods to handle class overlapping issues in various application domains

Category	Reference	Advantages	Disadvantages	Limitations
Under Sampling	Pattaramon et al. [114]	Improvement in minority visibility and accuracy	Excessive Elimination and information loss	Multi Class Dataset
	Rui et al. [115]	Improvement in F1-score and AUC	High time cost	Small Disjunct Class
	Debashree et al. [116]	Prevents from excessive elimination	Information loss	Noisy dataset
	Pattaramon et al. [100]	Improvement in minority visibility and accuracy, prevents from excessive elimination	Threshold Estimation	Used Only Soft Cluster
	Pattaramon et al. [117]	Minority Visibility	Excessive Elimination and Information Loss	Multi Class Dataset
	Soumya Goyal [71]	restricted excessive elimination and information loss	Not focused against minority performance	Used in software defect dataset
Over Sampling	Tao et al. [102]	Remove noise	Overfitting problem	Sub cluster division
	H Ibrahim [118]	Useful for multiclass, highly imbalanced and overlapped case	Clustering issue	Not useful for low imbalance and overlapped case
	Tao et al. [119]	Avoid generating noisy instances	No exact formula for K value estimation	Not Useful in multi-class dataset
Hybrid Sampling	Zhu et al. [120]	Strengthen the performance and obtained oversampling without introducing new samples	Optimal Value of K in KNN search	Optimization in evolutionary algorithm

networks, Naive Bayes and SVL trained with pre-processed data in the first stage. Pattaramon et al. [100] proposed an under-sampling method to capitalize the visibility of minority instances in the overlap region. The method uses soft clustering and elimination threshold criteria for overlapped instances. For optimality purposes only overlapped borderline instances are considered for elimination. Fuzzy C-means soft clustering is used for overlapped based undersampling of majority instances. Vuttipittayamongkol et al. [117] proposed an undersampling method that removes majority instances from overlap regions to improve the visibility of minority instances. The methods evaluate class membership degree (cmd) of minority and majority class instances using the Fuzzy C-means clustering technique and then cmd is used to eliminate majority class instances from the overlapping region. The majority of instances whose membership degree is less considered in the overlap region.

Over Sampling In oversampling-based method synthetic minority class instances regenerated subject to minimize overlap percent. Tao et al. [102] proposed SVDD and density peaks clustering oversampling (DPCO) technique for identifying optimal overlapped regions and regenerating synthetic instances to overcome the deficiency in the overlapped and imbalanced dataset. The SVDDDPCO finds dirty minority instances, eliminates them, and oversampled minority instances on the boundary. Ibrahim [118] proposed an outlier detection-based oversampling technique (ODBOT) for imbalanced dataset learning which reduces the risk of class overlapping. The ODBOT generates the synthetic instances of minority classes and later on, it detects outliers to reduce overlapping degrees in classes. The first step method determines the number of synthetic samples (NSS) minority class by the formula NSS= (P*M)/100 where M is the number of majority instances and P= (imbalance ratio-1)*100. Then in the second step, the method computes the dissimilarity relation between classes using a k-mean clustering algorithm. The algorithm finds k clusters of both majority and minority classes with a centroid of each. The third step method detects outliers in the minority class by the different relation between the cluster center of the majority and minority class. The summation of the minority cluster distance (SMCD) is computed as the difference between centers of minority and majority classes. The minimum SMCD means the minority class cluster contains outliers. Finally, the synthetic sample was generated to the boundary of minority instances of best clusters. Tao et al. [119] proposed SVDD based oversampling method for imbalance and overlapped domains. The SVDD is used to identify boundaries to avoid overlap then synthetic minority samples are generated using the weighted SMOTE technique. Maldonado et al. [121] proposed an oversampling method based on feature weighted to reduce overlap and imbalance. The method first selects relevant features based on the threshold value and then generates synthetic minority instances.

Hybrid Sampling In the hybrid, sampling-based method majority instances are removed from the overlapped region, and regenerates some minority class instances. Yuanwei et al. [120] proposed a hybrid method that removes the majority of instances from overlap regions that is less informative for the learning algorithm. The method performs under-sampling in the overlapped region and oversampling in non overlapped region. The method first detects overlapped regions and then finds out the optimal majority samples for elimination subjects to minimize the imbalance ratio, minimize the overlap ratio, and minimize the loss of original information.

Instance Overlap Based Methods: These methods handle the overlapping issue in two phases: In the first phase novel algorithm use split, entire data into the overlapped and non-overlapped region, and in the second step one of the following three approaches are used to reduce the negative effect of classification models. A support vector data description (SVDD) method was proposed in [122] which identified whether instances create overlapping or not. The SVDD finds the class domain boundary with minimum area by shrinking the radius of data space in this class. Once both class domains are identified by SVDD then common data instances belonging to both classes are put into overlapped regions. Xiong et al. [84] conducted a systematic study on existing methods of handling region-based class overlapping problems. Three schemes were proposed in past years' research which are discarding, merging, and separating overlapped regions. The separating-based methods give a higher performance in comparison to the remaining two, especially when using SVM, KNN, NB, and C4.5 classifiers. The summary of the instance overlap-based methods is shown in Table 3.

Re-sampling In these methods, under-sampling, oversampling, and hybrid sampling are performed on instances nearer to the borderline or in an entire overlapped region. Kumar et al. [1] proposed entropy and improved k-nearest neighbor based undersampling that eliminates only those majority class instances which having less potential informative. The overlap region identified by using improved kNN search and then eliminated majority instances if their entropy value is less than threshold entropy. Pattaramom et al. [3] proposed four methods to handle imbalance based on under-sampling by eliminating the majority instance from the overlapped region. All four methods are based on searching the target majority instance for elimination if there exists at least one minority instance within k nearest neighbor of the target instance. In the first method, the majority instance is eliminated without any consultation of the minority instance i.e. a majority instance is eliminated when its k nearest neighbor has a minority instance. In the second method, a majority instance eliminated when its k nearest neighbor has a minority instance, and further, it is also must be within the range of the k nearest neighbor of this minority instance. In the third method all majority instances are eliminated which is common in the KNN of two minority instances. In the fourth method, the elimination is further recursively applied to the result set which is to be eliminated in the third method. The proposed methods maximize the visibility of monitory instances in overlap regions and also minimize excessive elimination. It optimizes the trade-off between minimum information loss and maximum sensitivity. In these methods, there is the chance of excessive elimination of majority instances which decreases the accuracy. Only uniform distribution of instances has been considered in these methods. Estimating K-value in KNN is play an important role while identifying target instances. The higher k value can be used when minority density is lower than majority class density otherwise small k value is good for neighbor search. Sima and Hamid [123] proposed under-sampling (DBUS) and hybrid sampling (DBHS) two methods based on the density of instances. The DBUS deletes neighbors of higher density instances which reduces overlapping and balances data distribution while DBHS used a combination of under-sampling and oversampling for the same purpose.

Overlap Discarding Based Method In this method, the model ignores the overlapped data and learns from the remaining non-overlapped data, and later on, testing is performed on a model trained on this subset of data. The most common overlap discarding technique is SMOTE + Tomek Links used to handle problem [98]. Ronal et al. [101] proposed a model to evaluate the performance of clustering techniques to handle class overlapping. The result showed that soft clustering versions like FCM, rough c-mean, and their hybrid combination are powerful in comparison to other clustering techniques. FCM computes the membership of instances from each cluster center based on the distance similarity metric. Instances closer to the cluster center assigned high membership degrees from this cluster.

Overlap Merging Based Method In this method, the overlapped data instances are assigned a new class label and the original dataset includes one additional label named "overlap-label". Then after that, the existing model continues its learning phase, and two models are required for this scheme, one is for assigning a new label to instances of the overlapped region and the second is for the overlapping region. During the testing phase,

Table 3 Overview of instance overlap based methods to handle class overlapping issues in various application domains

Category	Reference	Advantages	Disadvantages	Limitations
Re-sampling	Kumar et al. [1]	Prevents from excessive elimination	Information loss, Estimation of entropy threshold	Multi Class Dataset
	Pattaramom et al. [3]	Prevents from excessive elimination	Information loss, Estimation of k	Multi Class Dataset
	Sima and Hamid [123]	Robust with noisy data and performs well for arbitrary shape data	Difficult estimating radius and min points parameters	Not appropriate when density deference is large
Separating	Sumana et al. [125]	Deals imbalance and overlap in high dimensional data, Efficient for both class	Optimal values of K in KNN search	Accuracy depends on clustering algorithm
	Shivani et al. [126]	Used efficient identification of overlap region	Minimized overfitting	Tested only limited overlap percentage
	Pasapitch et al. [127]	Give the best performance to classify minority data	Retaining efficiency to classify majority data	Linear kernel function deals only TPR
	Ronal et al. [101]	Used both hard and soft k means clustering	Used synthetic data	Not used real data

each target data is first tested by model one if it belongs to the overlap label then tested after it will be on the second model to determine actual its label.

Overlap Separating Based Method This method deals with overlapped and non-overlapped data separately for this purpose it requires two different learning models corresponding to each region. Then target data tests on individual models and determine whether it falls in the overlapped or non-overlapped region. Zian et al. [124] proposed a method that divides the dataset into overlapped and non-overlapped regions using k-mean clustering later on each region is trained and tested separating using two different classification models and finally all results are combined. Li et al. [106] proposed a hybrid-based method to handle imbalanced data with class overlap for fraud detection. The method is based on the divide and conquers approach. The dataset is divided into overlapping and non-overlapping subsets. Unlike other methods, the proposed method is not used any distance metric for obtaining overlapping regions. The method used an unsupervised anomaly detection model to learn the basic profile of the minority sample which belongs to fraud transactions. By using this learn profile an overlapped set can be formed which includes almost all minorities (fraud transaction) and some majority which is highly matched with the learning profile. The rest of the minority and majority samples form a non-overlapping set. During the conquer phase two subsets were tackled independently. Since the non-overlapping set has only a few minority samples in the non-overlapping region hence they can be classified as the majority one for simplicity. For overlapping sets, a powerful supervised classifier is applied for learning in the overlapped region. For assessment purposes method proposed dynamic weight entropy (DWE) to assess the quality of the overlapping and non-overlapping sets. The DWE is the multiplication of signal to noise ratio of minority class and entropy of overlapped subset.

Sumana et al. [125] proposed a model which initially divides the original dataset into overlapped and non-overlapped regions using the k mean clustering algorithm and then applied SVM optimization on each region separately to classify target data. Various data pre-processing applied before clustering such as handling missing values, min-max normalization, data balancing, removal of Tomek link, feature selection using random forest, and box plots to remove noise. Gupta and Gupta's [126] proposed a method to handle overlapping based on outlier instance identification and in the dataset. The proposed method measures the overlap degree by sequential using of the Nearest-Enemy-Ratio, Sub-Concept-Ratio, Likelihood-Ratio, and Soft-Margin-Ratio. Finally, the overlapped instances separated and said to noisy data and NB, C4.5, KNN, and SVM classifiers trained by non-overlapped data regions. Chujai et al. [127] proposed a cluster-based model which separates the dataset into the overlapped and non-overlapped regions and then both regions use to train two different classifiers for the prediction of unknown target data. Xiong et al. [103] proposed a method based on the Naive Bayes (NB) classification model in which NB is used to identify the overlapped region and further NB classification model trained on overlapped and non-overlapped region data separately. Liu et al. [128] proposed an anomaly detection model for the overlapped region in system log-based data. The model first detects the membership of an instance against both class label i.e. normal or abnormal and then use fuzzy KNN use to separate overlapped and non-overlapped region. The Adaboost-based ensemble learning is used to train the classification model and finally soft and hard voting methods are used to predict the class of unknown target data. Miguel et al. [129] proposed an n-dimensional generalized overlap function to determine the overlap degree in the imbalanced dataset. In the proposed function they generalized it from the multiclass data set to the binary class dataset. The difference is only in the number of sufficient boundary conditions, in n-dimensional it is more than the general overlap function.

3.1.2 Algorithm level methods

These methods do not change the class distribution of training data, they performed the learning process so that the importance of the minority class which is the class of interest in the learning domain, increases [38]. Most of the methods are modified based on the consideration of either class penalty or class weight or shifting decision threshold manage such a way the biasing towards majority is reduced. In a cost-sensitive method higher cost is assigned to the misclassification cost of the minority class [130, 131]. These methods are applicable on both instance level and also feature level selection and further sub categorized into cost sensitive and threshold.

Cost Sensitive Based Algorithms Cost-sensitive feature selection methods are proposed in [132] to reduce the effect of overlapping. Bo-Wen et al. [133] proposed density-based adaptive k nearest neighbor (DBANN) which handle both overlapping and imbalance problem together.

The DBANN uses an adaptive distance adjustment strategy to identify the most appropriate query neighbors. The method first divides the dataset into six partitions using the density-based clustering method which are minority noise samples, majority noise samples, minority samples in the overlapping cluster, majority samples in the overlapping cluster, minority samples in the non-overlapping cluster, and majority samples in the non-overlapping cluster. Now every six sets assigned a reliable coefficient to each training instance unlike KNN in other methods. The distance metric of each part is modified using local (within a partition) and global (entire dataset) distribution. Finally, the query neighbor is selected using a new distance metric. Saez et al. [146] proposed the use of one-vs-one (OVO) strategies to reduce overlapping degrees without any modification to the original data and the standard algorithm. The OVO decompose the multiclass problem into sets of binary class problems for example it divides C class problems into C(C-1)/2 binary sub-problems and each sub-problem deal with different classifiers separately. OVO can increase the separability between binary class problems, thus reducing the overlap degree and increasing the performance of classification models. Lee and Kim [139] proposed an overlap sensitive margin (OSM) model based on a fuzzy support machine and KNN to handle overlapped and imbalanced datasets. The OSM divides the entire dataset into hard and soft regions then regions classified using decision boundary 1NN and SVM. Using hard and soft regions it is easy to determine the closeness of the unknown dataset. Tung et al. [104] proposed an extended supervised dimensional reduction method more likely to principal component analysis to reduce overlap degree and increase the model performance on the overlapped and imbalanced dataset. The method advised learners to preserve the predictive power of smaller dimension subspace. It maximizes the inter-class covariance and minimizes inter-class covariance.

Threshold Feature Selection Based Algorithms These algorithms select a subset of key features instead of all for training and learning purposes. Although these methods are more challenging underclass overlapped in an imbalanced scenario in terms of data complexity but give good results in f-score and g-mean metrics. The penalty-based feature selection method is proposed in [142]. The feature subset selection method namely modified Relief that can handle both feature interactions and data overlapping problems. The method by assigning appropriate rank and penalty to each feature by their relevancy in data space and then the distance-based penalty discriminates against features whether they provide a clear boundary or not between two classes. Fatima et al. [143] proposed a feature selection method to minimize overlapped degrees to improve imbalanced learning in fraud detection applications. The method uses adaptive feature selection criteria to reduce overlap degree and then combined it with sampling strategies like no-sampling, SMOTE, and ADASYN. For feature selection, R-value uses to measure the overlap degree. The R-value is based on the average number of differences between the nearest neighbor and threshold nearest neighbor of each class. Using the R-value of both classes, R-augmented of predictor set of datasets computed which use to measure overlap degree in the dataset. As imbalance increases the augmented R-value also increases, therefore, the minority class has a larger weight in comparison to the majority class.

Feature Weight Based Algorithms These categories of algorithms assign weight to each feature according to their negative effects on the overlapped dataset. An adaptive method proposed in [144], which assigns a weight value to each feature according to the overlap degree measured using maximum Fisher's discriminant ratio (FDR). Datasets with a low value of FDR have a high degree of overlap. The F value is the maximum FDR valued feature computed by using (15). Xiaohui et al. [141] proposed a recursive features elimination method based on a support vector machine (SVM-RFE) to handle the overlapping problem in high dimensional data. The iteration of SVM-RFE ranks features based on weight and the bottom-ranked feature is eliminated. The weight of the feature determines by SVM-RFE-OA which combines the overlapping degree and accuracy of the samples. The overlapping degree of the sample xi is given by equation r(x_i)= different label(x_i)/k -OR(x_i). Where different label(x_i) samples have a different class label in k-nearest neighbors of a sample x_i and OR(x_i) is a non-homogeneous sample ratio in a dataset. The ratio r(x_i)>0 denotes sample ratio of different classes in neighbors of xi is greater than the heterogeneous data sample ratio in the original data set. For correct measure of overlapping degree is normalized Nr(x_i) form of r(x_i) need to be evaluate which is obtain by Nr(x_i)= r(x_i) / OR(x_i). The small average Nr(x) value of all samples represents high separability among the different class labels. The Nr(x) can represent descriptive information for discrimination. Shaukat et al. [147] proposed overlap sensitive artificial neural network (ANN) which handle both imbalance and class overlapping situation together with noisy class instances. The basic idea behind the method is weighting instances based on feature location before training to neurons. The advantage of these methods is to identify instances in the overlapping region instead of the entire region like other methods. For handling the overlap method apply weight to the instances concerning overlap level. The higher weight means less overlapping and the lower weight means more overlapping. For simplicity KNN(K=5) is used to evaluate propensity P=NN/5 where NN is the number of instances of the same class. The P=0 value represents an instance that lies in other class regions and is treated as outliers. The value P=1 represents instances surrounded by the same class instances. The value $0<P<1$ represents the overlapping of instances. The method uses a propensity range of outliers [0.0, 0.20]. The summary of the algorithm-based methods is shown in Table 4.

3.1.3 Other emerging methods

Unlike the re-sampling based method these need not change the class distribution of the original dataset but use random sampling with replacement to create n training sets of an equal number of instances of the dataset. The specific classification algorithm was applied to each training set and took majority voting as the final prediction of unknown target data.

Ensemble Based Methods These methods can be categorized as either homogeneous or heterogeneous depending on whether the classification models are the same or different for each training set during ensemble. Yuan et al. [148] proposed an ensemble method to handle overlap and imbalance in real datasets based on random forest trees called overlap and imbalanced sensitive random forest (OIS-RF). The OIS-RF derives a coefficient called hard

Table 4 Overview of algorithm based methods to handle class overlapping issues in various application domains

Category	Reference	Advantages	Disadvantages	Limitations
Threshold	Rubbo and Silva. [134]	Good performance in overlapped and highly imbalanced dataset	Setting high and low threshold	Not useful in multiclass dataset
	Zhang et al. [135]	Identify neighbors of unknown test instances	No formula for estimation K in KNN	Dataset validity missing
	Gu and Cheng. [136]	Provide better accuracy	Not optimal for all imbalanced class	Parameter optimization
	Afridi et al. [137]	Accurate estimation of overlap region	Not compatible in non overlap case	Clustering
Cost-Sensitive	Huaxiong et al. [138]	Reduces misclassification of minority	Overfitting	Not useful in large size data
	Lee and Kim [139]	Easy for learning when divided overlap into hard and soft regions	High generalization error	1NN used for extremely local search
	Ibomoiye et al. [140]	Force models to learn in minority region	Possibility of neglect majority instance	Performed only on few medical dataset
	Xiaohui et al. [141]	manage the trade-off between accuracy and overlapping	Typical feature selection criteria	Biological dataset
	Suravi et al. [142]	Handle features interaction and overlap together	Select only small subsets of best-performing features	Used distance-based measure
	Fatima et al. [143]	Used in different fields dataset	Can be used in higher dimensional dataset	Based on SVM and LR classifiers
	Saleh et al. [144]	Best performance by maintaining acceptable execution time	Difficulties in estimating weight	Optimal Weight
	Zhang et al. [145]	Used in multi-objective optimization	Useful for measurable feature cost only but in real application it is uncertain	Optimal Feature Selection

to learn (HTL). The HTL coefficient ensures the priority and importance of the instances at training time to the classifiers. The HTL depends upon overlap degree (OD) and imbalance ratio (IR). The OD(x) = (k1+1)/(k+1) where k1 is several instances of other classes in k-nearest neighbours of x. For instance x belongs to majority class HTL(x)= OD and HTL(x)=OD.IR for instance belongs to the minority class. This method provides higher precedence to the minority class instances that belong to the overlap region for training to the random tree. The algorithm gives better results in comparison to another algorithm in the context of the f-1 score, AUC, and precision for large and medium-sized datasets. Nve and Lynn [149] proposed KNN based under-sampling (K-US) method to filter instances from the overlapped region. The proposed model works in two stages i.e. data pre-processing and learning stages. The K-US resolve overlapped and imbalanced problems. In the learning stage, pre-processed data is used to train J48, KNN, Naive Bayes, and further ensemble with the AdaBoost strategy. Afridi et al. [137] proposed a three-way clustering approach to estimate the threshold which plays a major role to identify the overlapped class region. The threshold estimation is based on determining whether an instance belongs to a particular cluster, outside the cluster, or partially belongs to a cluster. The threshold value of an instance x is the ratio of the number of the neighbor of x that belongs to a particular class and the total number of neighbor objects. Yan et al. [150] proposed a method based on the local density of minority class instances in which the overlapping degree of a majority instance depends on the local density of surrounding minority instances and is then eliminated. The oversampling of minority instances is based on its local density to balance distribution.

3.1.4 Hybrid based methods

These methods combine both re-sampling and ensemble-based methods. Ensemble applied after reducing overlap degree by using any of the re-sampling methods i.e. under-sampling, oversampling, or hybrid sampling. These methods are time-consuming but useful where other performance metrics are desirable along with accuracy. Zian et al. [124] proposed a neighborhood under-sampling stacked ensemble (NUS-SE) to handle imbalanced and overlapping in the dataset. The NUS categorized data instances into safe, borderline, rare, and outlier. The borderline is overlapped instances and the outlier is noise instances. Unlike other under-sampling methods, NUS assigned instances weight according to neighborhood instances. The main idea behind NUS is to select higher-weight majority instances except for outliers for the training subset. The weight function of the instance is computed by the formula NUS(r) = exp(ar) where a is the exponential rate and r is the local neighborhood ratio of several minorities and the number of k nearest neighbors. After pre-processing using NUS then ensemble stacking is done which is the majority voting result of the heterogeneous classification model on different samples with replacement. The advantage of this method is to avoid the elimination of the informative majority class. Chen et al. [70] software defect prediction model which combines class overlapping reduction and ensemble learning to improve the prediction of defected software in imbalanced data. In the proposed method first neighbor cleaning rule (NCL) is employed to eliminate the majority of instances from the overlapping region then random under sampling-based ensemble learning is employed on the balanced dataset generated in the previous step. During the NCL step sampling eliminates non-defective instances to reduce overlap degree but in the ensemble, phase under-sampling eliminates non-defective instances to reduce the imbalance. Z. Li et al proposed a dynamic entropy-based hybrid approach for credit card fraud detection [106]. Wang et al. [41] proposed a hybrid method based on extreme-SMOTE and synchronous_Sample_LearningMethod for detection of financial distress. The training dataset used for this purpose was highly imbalanced and overlapped.

Hybrid of Feature and Instance-Based Method This method combines both feature selection and a re-sampling scheme. For key feature selection, the R-value measure computes the rank of features based on overlapped degrees proposed in [143]. The method proposed three feature selection algorithms RONS (Reduce Overlapping with No-sampling), ROS (Reduce Overlapping with SMOTE), and ROA (Reduce Overlapping with ADASYN), which are derived throughout sparse feature selection to minimize the overlapping. Rubbo and Silva [134] proposed filter-based instance selection (FIS) method which uses self-organizing maps neural network (SOM) and entropy theory of information. The SOM learns from the training dataset after that training instances are mapped with the closest blueprint of the SOM neurons. Further training datasets are filtered by satisfying the minimum threshold in the majority class filter (high filtered instance selection (HFIS)), minority class filter (low filter instance selection (LFIS)), and both class filters (both filter instance selection (BFIS)). These filtered instances are depending on the entropy of neurons. If the entropy is greater than or equal to the high threshold value, then overlapped instances which is less informative are removed (HFIS).

If entropy is less than or equal to a low threshold then lower probability class instances removed by ensuring removed instances do not with majority class can smooth the class boundary(LFIS). The BFIS combines both HIF and LFIS and border overlapped instances are removed. After the filtering process is done the selected dataset is now used for training the model used for target prediction. The disadvantage of the method is to set threshold parameters. The summary of the other emerging based methods is shown in Table 5.

$$f = \frac{{\mu_1}^2 - {\mu_2}^2}{\sigma_1^2 + \sigma_2^2} \tag{15}$$

where μ_1, μ_2,, σ_1^2 , σ_2^2 are the means and variances of class1 and class2.

3.2 Deep learning based review

Deep learning is an associate field of machine learning (ML) which includes artificial neural networks (ANN). The ANN is derived from interconnected biological neurons or nodes [38]. The main difference between ML and DL is the types of data used for learning. The ANN is made of a user-defined input layer, a hidden layer (calculate the result of the input layer), and an output layer as the result of the final calculation. The weighted connecting line joins nodes between adjacent layers. The DL is made of deep neural networks i.e. three or more hidden layers. Each neuron transfers its input into single output by using a nonlinear activation function in a feed-forward manner. A fully connected neural network has one hidden layer called a multi-layer perceptron (MLP). An MLP is simplest form of DL [61, 152–154]. Jeong et al. [57] presented a technical review on GAN-based methods and focused on major GAN architectures used in medical image augmentation and balancing for improving classification and segmentation tasks. A fully connected three-layer feed-forward neural network is shown in Fig. 7.

Aksher et al. [40] proposed CNN based model for sock price prediction. The model handle imbalance and feature selection problem. They used 2D deep neural network with a new rule-based labeling algorithm and feature selection model. The approach gives out-performance but it is limited to only stock trade prediction. Zhao et al.[30] proposed normalized CNN-based models for rolling bearing machinery fault detection and diagnosis. The model used

Table 5 Overview of other emerging methods to handle class overlapping issues in various application domains

Category	Reference	Advantages	Disadvantages	Limitations
Ensemble	Bo-Wen et al. [148]	Prevents Information Loss	Optimal Parameter Setting	Used only CART
	Chunbo et al. [128]	Use a combination of fuzzy set and KNN for overlap detection	possibility of error due to higher fuzziness	Not useful in high dimensional data
	Zian et al. [124]	Adopted meta learner in stacked ensemble	High complexity	Selection of Meta Learner
	Lin Chen et al. [70]	Removes noisy instances from overlapped region	Ensemble random undersampling may suffer excessive elimination and information loss	Limited only software defect problem domain
	Anandarup et. al [151]	Useful in the multiclass dataset and adopted dynamic ensembling	Time cost high due to pre-processing	May not cope with multiclass and different reprocessing methods
	Everlandio et al. [105]	Optimized samples used in different base classifiers	Higher time cost	Generate samples with help of induced models
Hybrid	Z. Li et al. [106]	Designed for a trade-off between minority outliers and IR in overlapped region	Overfitting	Parameter selection methods in decision-making
	Wang et al. [41]	Extreme_SMOTE and synchronous sampling method	One F-1 score considered	Larger time cost due to threshold search by classifier
Local Density	Yan et al. [150]	Utilize both borderline and safe area phenomena	High complexity	Not useful in multiclass dataset

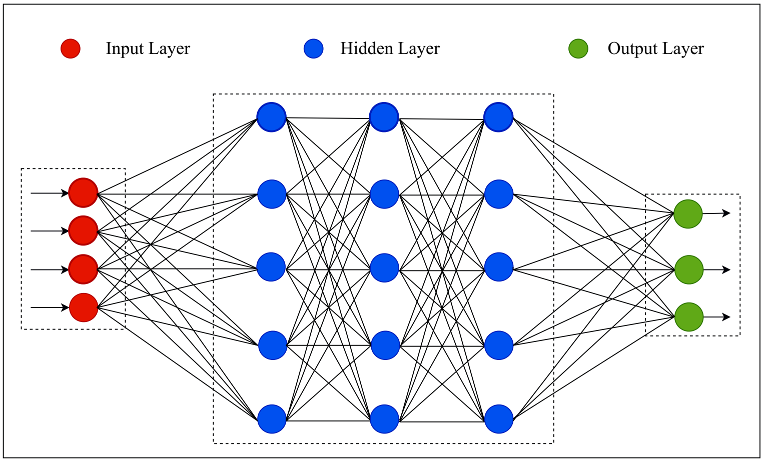

Fig. 7 Deep Learning Neural Network with multiple hidden layers. The left most layer is used to supply input and right most layer is used to represents output. Three intermediate hidden layers are used to perform non-linear transformations on supplied input to generate output

the CNN layer for mapping the feature through a series of convolutional operations. The model gives good accuracy in imbalanced conditions. Wang et al. [155] proposed CNN to identify particular emitters with a received electromagnetic signal. The model tackles the imbalance problem in both gain and phase of signal context and improved the performance over the traditional feature-based approach. Chen et al. [156] proposed an ensemble-based deep learning (DL) convolutional neural network (CNN) method to handle imbalanced data. They designed an improved loss function to rectify biasing of learning models towards the majority class. The loss function enforces CNN hidden layers and other associated functions which reduce misclassified samples in lagging layers. The method initiates adjacent layers to create miscellaneous actions and re-fix errors in former layers according to a batch-wise manner. The major problem with the method is not combining advanced loss functions and novel data generation for better learning capability in high-quality data.

Yan et al. [157] proposed an extended DL method for the unbalanced multimedia dataset. The method integrates bootstrapping, DL and CNN. The method fed low-level features to CNN and justified feasibility in achieving out performance with reduced time complexity. Akil et al. [59] presented a new deep CNN for automated low and high-grade segmentation of brain tumors to maximize feature extraction from MRI images. The proposed method resolves two major issues class imbalance and overlapped image patches. For handling imbalance issues they use equal sample images and examine loss function results. Antom Bohm et al. [158] presented an approach using full CNN for image object detection and segmentation. The approach translates overlapped image objects into non-overlapped equivalents. Banarjee et al. [159] proposed two approaches based on CNN and RNN for information synthesizing computed tomography of clinical data. The model presented a report in word, sentence, and document levels. It outperforms in f1 score value which is desirable in the highly imbalanced dataset. Gao et al. [160] proposed a method based on three layers of CNN with two dense layers. CNN uses thirty-two one-dimensional kernel filters in each of the three layers having sizes three, twenty-two, and twenty-one respectively. The layers process one-dimensional raw data and extract useful features which supply dense layers for classification. The problem with this method is to use simple imbalance solutions.

Rai and Chatarjee [161] proposed a hybrid CNN_LSTM and ensemble-based model which automatically detects the cordial attack in ECG data. SMOTE oversampling technique was used to balance the dataset and better performance in minority class accuracy. It is used in clinical diagnosis and need not to required feature extraction and other preprocessing technique. Gupta et al. [22] proposed a model to tackle class imbalance based on LSTM and improved one-vs-one in frequent and infrequent intrusion detection in the network. The model achieved a higher detection rate with reduced computational time. Gao et al. proposed LSTM_RNN for arrhythmia detection in ECG imbalanced data [162]. The LNN is used to retain valuable information. To handle skewness in data SMOTE, random under-sampling and distribution-based methods were used. Tran et al. [163] proposed LSTM based framework to handle an imbalance in the multiclass dataset. The approach uses to detect botnets in domain generative algorithms. It achieves a 7% higher micro average in precision, and recall as compared to main LSTM and other cost-sensitive based methods.

Dong et al. [164] developed a novel DL model for handling imbalanced data. The model is based on incremental minority batch rectification using hard data sample mining at learning time. The model degrades the dominant effect of the majority by identifying the boundary of sparse minority samples in a batch-wise manner. The method includes the rectification loss function of a class that deployed deep neural structural design. The problem with the model is not considering the majority class and focusing only on the minority for making a cost-effective class rectification loss (CRL) function. Lin et al. [165] proposed a model based on reinforcement DL to formulate the problem by sequential decision-making process using deep quality of state-action function. In this method, the agent acts awarding or penalizing the learning model on the input sample. The model becomes more sensitive towards minorities by rewarding them more in comparison to the majority sample. The major problem with this model is to design a novel reward function and learning algorithm for a multiclass dataset. Yuan et al. [17] introduced a regularized framework for ensemble DL to tackle the multiclass imbalance problem. The method employs regularization for penalizing classifiers in the misclassification stage which were correctly classified in the previous phase. During the training phase repeatedly misclassified treated as a hard sample and extraordinary focus is required of the learner. The drawback of the method is used only in stratified sampling cases and has higher computational time complexity.

Wang et al. [166] proposed a deep neural network (DNN) for handling imbalanced problems. The DNN mainly focus on CRL to achieve optimal parameter for iteratively minimizing training errors. DNN uses the mean squared false error (MSFE) loss function to confine equally majority and minority class errors. It is inspired by the concept of true positive and negative rates. Limitation of DNN to explore effective MSFE loss functions for different NN architecture. Stoyanav et al. [58] proposed dice loss function (DLF) to examine the sensitivity of loss function for learning and different imbalance ratio of labels in 2D and 3D segmentation tasks. They also proposed maintaining the class imbalance property of the DLF. Zhang et al. [167] proposed a deep belief network based on evolutionary cost sensitivity (ECS) for imbalanced data. It optimized the misclassification cost of training data based on the G-mean objective function. The proposed model first optimizes misclassification error cost and then applies DBN. Dablain et al. [168] proposed DeepSMOTE which contain three components: encoder/ decoder, SMOTE, loss function. It generates high-quality and informative artificial images which are right for visual examination. The problem with the method does not apply to graph and text-based data. Andrei et al. [169] demonstrated how DL techniques can be utilized to detect overlapped speech. For this purpose, they use 10 different speakers to produce overlapped speech and analyze how DL techniques are helpful in this aspect.

Alia et al. proposed [170] hybrid DL and visual framework for detecting the behavior of pedestrians. The two main components of the work are extraction of motion information and annotation of pushing patches. The combination of CNN + RNN was used for calculating the density of the crowd. The limitation of the work is, that it cannot be used in real applications but it depends on recorded video. It will not work in case of recording with a live-moving camera. Wang et al proposed [171] DL generative model which is vigorous to imbalanced classifications. The new synthetic instance is generated with the cooperation of the Gaussian distribution property. The work is limited to the integration loss while producing a new synthetic instance. Liu et al. [32] proposed a model based on GAN with feature enhancement (FE) used for fault detection purposes in mechanical rolling bearing. The GAN extends imbalance data and FE collects faulty features. The model is not applicable in semi-supervised learning examples. The training process requires a lot of resources hence computational complexity is high. Yue et al. proposed [172] modified GAN to balance and augment Egyptian characters written on bone in the form of hieroglyphs. The dynamically selected augmented data then use to train the DL classifier. Using this concept they proposed a novel model for character recognition. It is useful only for larger datasets. Liu et al. [173] presented an overview of DL-based industrial applications by enhancing samples. They also discussed explainable DL methods. Arun et al. proposed [174] a novel model based on LSTN + RNN to predict country-wise cumulative COVID-19 infected cases. The LSTN, RNN with the combination of gated recurrent unit performs nonuniform is a major problem with this model even though it depends on regionalism data. Liu et al. [24] proposed a novel data synthesis method based on GAN for enhancing deep features. Later on, it tested imbalanced data for fault diagnosis in the rolling bearing machine. The time complexity is high due to the adversarial training strategy. Ding et al. [19] designed a model based on GAN for intrusion detection in a network. Hybrid sampling is used to reduce the imbalance ratio. The KNN and GAN are used for under and oversampling respectively. The model has not considered the effect on the majority while generating a minority sample. It is unable to solve the overlapping problem. The summarized issues and limitations in deep learning based models are given in Table 6.

Table 7 shows top deep learning neural networks along with brief descriptions and application areas.

3.3 Big data based review

Yin et al. [7] proposed an approach for efficient and safe Tunnel-Boring-Machine operation on imbalanced big data. In this approach, they first preprocessed training data using a mining technique and then used the adaptive-synthesis (ADASYN) algorithm to reduce the imbalance. finally, hybrid ensemble learning is used to identify the geological rock class. khattak et al. proposed deep learning based model for detecting theft activity by illegal pattern in electricity consumption. The challenging task in such bi data is class imbalance which was resolved by using adaptive synthetic and Tomeklinks method [89]. Sripriya et al. [88] proposed model for handling imbalance issue in drug discovery big data environment. to handle class imbalance issue, synthetic instances generated using SMOTE then combination of Gradient boost, logistic regression and k-nearest neighbor techniques used to protein selection for discovery of relevant drugs. Javaid et al. [8] proposed a method for electricity theft detection using LSTM+CNN on big data. The imbalance is handled by the ADASYN approach where theft behavioral instances are in minority and class of interest. Jonson et al. [204] conducted experimental study based on random undersampling, random oversampling and its combi-

Table 6 Issues and challenges in deep learning based models used to handle imbalance and overlap problems

Approach	Limitations	Reference
CNN	Accuracy testing could not guarantee reliability.	[175]
DL and ML	Alone performance metric is not sufficient to tackle the real-world task. For industrial applications, it is like a black box for workers.	[176]
Neural Network	lack of explaining ability are strong barriers to future development.	[177]
CNN + RNN	Cannot used in real-world application	[170]
DL-based generative	Limited to the integrating loss while producing new synthetic instance	[171]
Modified GAN	Not Applicable for smaller dataset	[172]
GAN	Not useful for Semi-supervised examples	[32]
Deep + SMOTE	Not application in text and graph data	[168]
DNN	Limited to the exploration of the loss function	[166]
Reinforcement + DL	Difficult to design reward and penalty actions.	[165]
CNN	Limited to the particular application only.	[40]
Ensemble + CNN	The major problem with the method is not combining advanced loss function, novel data generation for better learning capability in high-quality data	[156]
3-layer CNN	Not suitable for highly imbalanced data cases.	[160]
CNN+LSTN	Application dependent used only in clinical diagnosis	[161, 178]
RNN+LSTN	Performance varies over regional data. Also, depending on the combination of a gated recurrent unit with both RNN and CNN	[174]
GAN	Time complexity is very high	[24]
GAN + KNN	Model unable to solve overlap problem. It also ignores the effect on the majority while generating a minority sample.	[19]

nation to handling imbalance problem in big data domain and concluded that oversampling performs better than undersampling. Wang et al. [11] proposed an imbalanced and big data framework for extreme ML in distributed and parallel computing environments. Sleeman et al. [9] proposed a framework for a multiclass imbalanced big data environment based on novel SMOTE to overcome data distribution limitations. The framework is augmented through informative resampling and novel-portioning-SMOTE on spark nodes for a big volume of data. Maurya et al. [10] presented a framework for imbalance and big data domain. The method outperforms in G-mean,f1-score, and error rate on several big volume datasets as compared to state-of-the-art methods. For maximizing the classification performance metrics convex loss function was introduced. Justin et al. conducted a study for evaluating the use of deep learning and resampling on imbalanced big data in various domains such as fraud detection and medicare [12]. The study concluded that random sampling is the most preferred method to handle imbalanced big data issues. Zhai et al. proposed an under-sampling and ensemble-based method for the classification of larger volume datasets. It used the MapReduce approach to translate big data clusters into subsets to maintain data distribution adaptive manner. Yan et al. [205] proposed a framework based on borderline margin loss by separating minority and majority samples in the overlapped region for a bigger volume dataset. Major big data issues is shown in Table 8.

Table 7 Methods and neural network for handling class overlap and class imbalance in deep learning environment

Deep Learning Network	Description	Usage	Reference
Convolutional Neural Networks (CNNs)	It Has multiple hidden layers in a neural network.	It is used for the detection and processing of image objects.	[30, 40, 59, 155–160]
Long Short-Term Memory Networks (LSTMs)	These Networks Recalling its past prediction for learning.	It is used in time series prediction, speech recognition, and music composition.	[22, 161, 174]
Recurrent Neural Networks (RNNs)	It Has directed cyclic connections as feedback.	It is used in time series, image caption, handwritten document recognition, and machine translation.	[90, 159, 174, 179]
Generative Adversarial Networks (GANs)	create new data instances for resembling the training dataset.	It is used to generate realistic images, cartoons, photographs of human faces, and 3D objects.	[19, 24–26, 180–184, 184]
Radial Basis Function Networks (RBFNs)	It is special types of feed-forward neural networks using radial basis activation function.	It is used for classification, regression, and time series analysis.	[185–188]
Multi-layer Perceptron (MLPs)	These Have fully connected an input layer, a hidden and an output layer.	It is used to make speech and image recognition, and machine translation software.	[18, 189–193]
Self Organizing Maps (SOMs)	SOMs use data visualization concepts.	It is used to create user understandable high dimensional data.	[20, 27, 194–196]
Deep Belief Networks (DBNs)	DBNs are generative models consisting of multiple layers of stochastic, hidden layers of a binary variable.	It is used for image and recognition, and capturing motion data.	[28, 167, 197, 198]
Restricted Boltzmann Machines(RBMs)	It is stochastic neural networks that learn from a probability distribution function of inputs.	It is used for classification and regression, dimensionality reduction, feature filtering, and learning.	[14, 15, 21, 199, 200]
Autoencoders	These are trained neural networks that replicate the data from the input layer to the output layer.	It is used in pharmaceuticals, prediction of popularity, and image processing.	[29, 107, 201–203]

Table 8 Issues and challenges in imbalance and overlap method in big data domain

Approach	Issues	Reference
ADASYN and hybrid ensemble	Regenerated synthetic instance may create overlap.	[7]
ADASYN + TomekLink	Regenerated synthetic instance may less informative and Tomelinks may suffer from information loss.	[89]
SMOTE and LR+KNN+GB	Redundant information may be generated by SMOTE.	[88]
ADASYN and LSTM+CNN	Performance depend on the selection of hyperparameter.	[8]
Map Reduce and extreme ML	Not works on complex ML algorithm.	[11]
Informative resampling and SMOTE	Results are strongly data dependent.	[9]
Borderline and overlap separation	Not worked in the entire overlapped region.	[205]

4 Comparative analysis with existing survey

This section describes a detailed comparative analysis of our survey with existing surveys in overlapped and imbalanced domains. The presented analysis is based on different characteristics whether these have been included in the survey or not. Class imbalance, class overlap, categorization of handling methods, advantages and disadvantages of the methods, limitations of the methods, important findings and research gaps, and performance metrics with highlighted outperformed one by handling method. The analysis includes surveys between the years 2016 to 2022 in the context of deep learning, machine learning, and big data environment. Table 9 provides a detailed comparative analysis of this survey with existing surveys.

Branco et al. [81] grouped existing methods to handle imbalanced and overlapped datasets into data pre-processing, special-purpose learning method, prediction post-processing, and hybrid methods. The data pre-processing approaches include methods that change class distribution in the imbalanced dataset to make it balanced. These methods are further grouped into resampling, active learning, and weighting data space. The advantages of these methods are: compatible with any existing learning tools and model classification models are more interpretable to reduce bias. The special-purpose learning method modifies the existing algorithm to make it compatible with imbalanced and overlapped datasets. The main drawback of this approach is the restriction of choice, changing of target loss functions, and prerequisite deep knowledge of algorithms. The prediction post-pre-processing approaches use the original dataset and the learning algorithm only manipulates the predictions of models as per preference to user and imbalance.

These methods were further regrouped into the threshold and cost-sensitive post-pre-processing method. The hybrid methods are obtained by combining resampling and special-purpose learning methods.

Neelam et al. [82] grouped existing solutions to handle imbalanced and overlapped datasets into data-level methods, algorithm-level methods, ensemble and hybrid methods, and other different methods. The data level methods resample the dataset to reduce the negative impact on the classification model. The algorithm-level methods modify the existing models or

Table 9 Comparative analysis of the existing and current survey

Year	Survey	Focusing Area	Summarized Report in the Survey							
			Imbalance	Overlap	Category	Advantages	Disadvantages	Limitations	Research Gap	Performance
2016	[81]	Addresses major challenges, description of approaches, comparative study, and analysis of the existing methods.	✓	×	✓	✓	✓	×	✓	×
2018	[82]	Summary of methods with their advantages and disadvantages, challenges, and performance analysis.	✓	×	✓	×	×	×	✓	×
2018	[83]	In-depth comparative analysis of methods and their grouping based on the technique used in the method.	✓	×	✓	✓	✓	✓	✓	×
2019	[38]	Implementation and experimental result details of methods handling imbalancement in deep learning based on neural networks approach.	✓	×	✓	×	×	×	✓	×
2021	[2]	A typical Review of existing methods and their categorization based on approaches used in the method.	✓	✓	✓	×	×	×	✓	×

Table 9 continued

Year	Survey	Focusing Area	Summarized Report in the Survey							
			Imbalance	Overlap	Category	Advantages	Disadvantages	Limitations	Research Gap	Performance
2022	[154]	Provided a comprehensive study about deep learning-based segmentation techniques in brain tumors.	✓	×	×	×	×	×	✓	✓
2022	[79]	Reviewed jointly both imbalance and overlap effects on the performance of the classifier. Discussed various overlapping region formulation measurements.	✓	✓	×	×	×	✓	✓	✓
	Current	Category wise summary, issues resolved, limitations, performance summary report, highlighted outperformance metrics of the methods, and listed research gaps for future directions.	✓	✓	✓	✓	✓	✓	✓	✓

The comparison based on major challenges addressed in survey like data intrinsic characteristics, categorization of the existing methods, pros, cons,limitation, used performance metrics and research gaps design a new algorithm which sensitive to imbalanced and overlapped datasets. The ensemble and hybrid method uses multiple samples with replacement of dataset and each sample trained either single model independently and behaves like multiple models and final decision of testing data taken by the averaging result of the model trained by different samples. The other methods are based on feature selection, instance selection, and principal component analysis by using different complexity measure functions.

Kaur et al. [83] conducted a review on imbalanced dataset challenges in machine learning applications. The authors presented reviews on methods based on data pre-processing, algorithm, and hybrid approach. The review search criteria are based on parameters: definition, nature, challenges, algorithms, evaluation metrics, and domain area of the problem and issue.

Johnson and Khoshgoftaar [38] categorized existing methods to deal with class imbalanced datasets for learning algorithms: data-level methods, algorithm-level methods, and hybrid methods. Data level methods effort to shrink the intensity of imbalance all through various data resampling methods. Algorithm-level methods for handling class imbalance, usually implemented with a weight or cost representation, comprising modifying the fundamental learner or its amount produced to decrease biasing to the majority instance. At last, hybrid systems purposefully merge both sampling and algorithm-based methods.

Pattaramom et al. [2] presented a technical review on existing approaches to handle class imbalance and class overlapping and its negative effects on the performance of classification models. The experimental results showed that increases in the overlap percentage allow decreases in the performance of models but in the case of class imbalance meant not always have a performance effect. Both overlap and imbalance decrease sensitivity but when the overlap percentage is small, changes in imbalance do not show any impact on model performance however at a higher overlap percentage, degradation of sensitivity is due to class imbalance meant. This review shows that the class overlapping effect degrades the sensitivity more in comparison to the imbalance. Xiong et al. [84] conducted a systematic study on the class overlapping effects and showed that overlapping-based methods are more powerful in comparison to other methods. In this method entire data is divided into the overlapped and non-overlapped regions by using a well-known classification algorithm and then overlapped region issue is resolved using any one of the following strategies: Discarding, Merging and Separating overlapped regions in which separating strategy is good and shows better improvement in performance metrics.

Santos et al. [79] presented a deft review on the combined effect of class overlap and imbalance on the performance classifiers. In this review, they specially focus on different measurements of overlap and analyze its formulation on varieties of datasets. Finally, the review summarized major shortcomings such as (i) there is no exact formulation for measurement of overlap. (ii) The studies represented overlap according to their methods which are an inconvenience for performance measures on a single platform. (iii) class overlap degree has not considered any other intrinsic or extrinsic characteristics except imbalance.

5 Performance comparison and research gap

This section presents a performance metrics summary of various existing methods with highlights of one in which the method outperformed compared to the state-of-the-art methods. And the second part of this section we summarized the important finding and research gaps in existing methods and approaches.

5.1 Performance metric summary

This section presents performance metrics of standard machine learning algorithms on dataset in which class overlapping and imbalance problems tackled by individual overlap-based method. Table 10 illustrates a brief overview of the performance metric summary of proposed methods in the literature, ✓ indicates that method has been evaluated in respective performance metric while dealing class overlapping issue. It is based on experimental summary and comparative analysis result provided in the literature of respective method.

5.2 Research gaps

Many methods had been proposed by the researcher in past years to handle class overlap effects on the performance of classification models in imbalanced datasets. The following most common research gaps are found in our survey:

- In previous approaches, researchers were concerned with only specific performance metrics and applications. For experimental purpose both real and synthetic was considered but applicable only to those datasets which satisfy normal probability distribution. Most of the real datasets are in which instances are not distributed normally.
- In most of the instance-based approaches, the challenging task is to divide the whole dataset into overlapped and non-overlapped regions. There is no such global method that can be used for this purpose.
- Due to the lacking of appropriate mathematical characterization to identify instances in the overlapping region, it is quite difficult to find out arbitrary shaped overlapped regions in a real-world dataset. Past research only assumes spherical or rectangular shape overlap regions.
- The existing approaches used KNN search to identify instances in the overlap region. Estimating the value of k is still challenging for researchers. Most of the methods used random value of k wherever few approaches estimated k value concerning data size and imbalance but could be used only in uniform dada density distribution scenario.
- Instance-based approach dealing overlap effect by eliminating majority instances which causes information loss due to excessive elimination.
- The popular neural network-based techniques in deep learning have high computational complexity and needed larger size data for training and learning purposes which may be unavailable.
- The algorithm-level approaches are based on two hyperparameter thresholds and misclassification cost estimations. For evaluating cost majority of these methods have been used SVM for initially misclassified instances and later on cost-sensitive based new algorithm designed on the modified existing one. Estimating optimal parameters is quite complex.
- Some new emerging methods like the local density-based approach are used in the latest trends which is not feasible in non uniform data distribution cases.

6 Conclusion and future scope

A systematic in-depth discussion throughout this study is conducted which deals class overlap issues in extreme imbalance real-world application domains. Included domains are health, security, financial, bio-informatics, data mining, machinery, and software fault, prediction

Table 10 Performance metric summary (Bold ✓ indicates that overlapping method out performed in the respective performance measure metrics by the standard machine learning classifiers)

Reference	Performance Metrics									
	Sensitivity	Specificity	Precision	F-score	G-mean	Balanced Acc	Accuracy	AUC	MCC	Comparative References
Kumar et al. [1]	✓		✓	✓	✓					[3]
Pattaramon et al. [3]	✓		✓	✓	✓					[117, 206]
Pattaramon et al. [114]	✓	✓		✓	✓		✓			[206]
Rui et al. [115]				✓				✓		[207, 208]
Pattaramon et al. [117]	✓					✓				[206]
Debashree et al. [116]	✓	✓	✓	✓			✓	✓	✓	[209, 210]
Pattaramon et al. [100]	✓	✓		✓	✓					[52, 117, 206]
Xinmin et al. [102]				✓	✓			✓		[84, 211–213]
H IBRAHIM [118]				✓	✓			✓		[214–216]
Zhu et al. [120]				✓	✓			✓		[217, 218]
Rubbo et al. [133]				✓	✓					[53, 117, 219]
Zhang et al. [135]							✓			[220]
Afridi et al. [137]	✓		✓	✓			✓			[221, 222]
Shaukat et al. [147]				✓	✓		✓			[223]
Lee and Kim [139]				✓	✓					[224–226]
Yuping et al. [136]	✓	✓		✓			✓			NA

Table 10 continued

Reference	Performance Metrics									
	Sensitivity	Specificity	Precision	F-score	G-mean	Balanced Acc	Accuracy	AUC	MCC	Comparative References
Huaxiong et al. [138]							✓			[227–229]
Ibomoiye et al. [140]	✓		✓	✓			✓	✓		[13, 230]
Bo-Wen et al. [148]			✓	✓				✓		[231, 232]
Seng Zian et al. [124]								✓		[233, 234]
Lin Chen et al. [70]	✓	✓			✓			✓		[235, 236]
Chunbo et al. [128]	✓		✓	✓			✓		✓	[23, 237]
Everlandio et al. [105]						✓				[54, 238]
Xiaohui et al. [141]	✓	✓					✓			[16, 239]
Saleh et al. [144]				✓						[240]
Suravi et al. [142]							✓			[67, 68, 241]
Zhang et al. [145]							✓			[242, 243]
Yufei Xia et al. [131]				✓				✓		[244]
Sumana et al. [125]	✓	✓	✓	✓			✓			NA
Shivani et al. [126]	✓		✓	✓			✓			NA
Pasapitch et al. [127]	✓	✓		✓	✓					[245]

computer vision and image processing, anomaly and spam detection, etc. where training data are highly imbalanced and overlapped. In these domain high accuracy of unknown data prediction for the minority class instances are expected. It is only possible when training data intrinsic characteristics will be rectified or designed such a robust machine learning algorithm for these datasets. Here in this work, detailed descriptions of existing class overlap handling strategies along with their compatibility, limitations, and major research gaps are presented in category-wise tabular format. Major challenging things among all existing approaches are the lack of a common global strategy for class overlap characterization, techniques to resolve issues, and attaining a satisfactory level of efficiency in desired performance measure metrics in the specific domain.

The current presented a detailed review and analysis of the negative effects of class overlap on classification models in imbalanced data domains. This paper also provides a comprehensive discussion on the current methods in the field of machine learning, broad classification, and comparative analysis of this survey with existing recent years surveys. Finally, study also showed that deep learning and machine learning algorithms have continuously drawn the attention of researchers in small, medium, and big data domains. The reason behind it they all have the capability of automated learning and producing feasible results. Although many algorithms have been proposed in imbalanced datasets very few of them focus on class overlap intrinsic characteristics of the small, medium, and big data domains. And these methods produce optimal results but popular neural network-based techniques in deep learning have high computational complexity and needed larger size data for training and learning purposes which may be unavailable in many domains like fraud detection, and health sectors. This may be one future direction for the researchers.

As we know that class imbalance and overlap problem could not be avoided in real datasets due to limitations in the data preparation phase and natural data intrinsic characteristics. To overcome these data issues existing approaches are implemented in two phases i.e. identification of overlapping regions and then reducing it at certain degree level. The approaches used for identifying overlapped regions in the state-of-the-art methods are according to author convenience and feasibility because there is no such global mathematical class overlap characterization formulated yet. Therefore it may be another future dimension for the researchers. Many data level-based methods used the KNN approach to identify instances in the overlapped region. Estimating optimal K may be another future direction while using such an approach to tackle the problem. Existing approaches designed for uniformly probability data distribution scenario which is not true in many current reals-worlds applications dataset, since almost of them follow nonuniformly probability data distribution, thus it may be another future direction for the researchers. The algorithm-level approaches are based on two hyperparameter thresholds and misclassification cost estimations. For evaluating cost majority of these methods have been used SVM for initially misclassified instances and later on cost-sensitive based new algorithm designed on the modified existing one. Estimating optimal parameters may be another future direction. Some new emerging methods like the local density-based approach are used in the latest trends which is feasible in uniform data distribution cases. Estimating the local density of instances in nonuniformly data distribution cases may be another future direction.

Funding This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability statement Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflicts of interests/Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Kumar A, Singh D, Yadav RS (2023) Entropy and improved k-nearest neighbor search based undersampling (ENU) method to handle class overlap in imbalanced datasets. Concurr Comput Pract Exp e7894
2. Vuttipittayamongkol P, Elyan E, Petrovski A (2021) On the class overlap problem in imbalanced data classification. Knowl-Based Syst 212:106631
3. Vuttipittayamongkol P, Elyan E (2020) Neighbourhood-based undersampling approach for handling imbalanced and overlapped data. Inf Sci 509:47–70
4. Bilal M, Maqsood M, Yasmin S, Ul Hasan N, Rho S (2022) A transfer learning-based efficient spatiotemporal human action recognition framework for long and overlapping action classes. J Supercomput 78(2):2873–2908
5. Ghosh K, Bellinger C, Corizzo R, Krawczyk B, Japkowicz N (2021) On the combined effect of class imbalance and concept complexity in deep learning. In: 2021 IEEE international conference on big data (big data), pp 4859–4868
6. Zhai J, Wang M, Zhang S (2022) Binary imbalanced big data classification based on fuzzy data reduction and classifier fusion. Soft Comput 26(6):2781–2792
7. Yin X, Liu Q, Huang X, Pan Y (2022) Perception model of surrounding rock geological conditions based on TBM operational big data and combined unsupervised-supervised learning. Tunn Undergr Space Technol 120:104285
8. Javaid N, Jan N, Umar Javed M (2021) An adaptive synthesis to handle imbalanced big data with deep siamese network for electricity theft detection in smart grids. J Parallel Distrib Comput 153:44–52
9. William C, Sleeman IV, Krawczyk B (2021) Multi-class imbalanced big data classification on spark. Knowl Based Syst 212:106598
10. Maurya CK, Toshniwal D, Venkoparao GV (2016) Online sparse class imbalance learning on big data. Neurocomputing 216:250–260
11. Wang Z, Xin J, Yang H, Tian S, Yu G, Xu C, Yao Y (2017) Distributed and weighted extreme learning machine for imbalanced big data learning. Tsinghua Sci Technol 22(2):160–173
12. Johnson JM, Khoshgoftaar TM (2019) Deep learning and data sampling with imbalanced big data. In: 2019 IEEE 20th international conference on information reuse and integration for data science (IRI), pp 175–183
13. Chatrati SP, Hossain G, Goyal A, Bhan A, Bhattacharya S, Gaurav D, Tiwari SM (2020) Smart home health monitoring system for predicting type 2 diabetes and hypertension. J King Saud Univ-Comput Inf Sci
14. Liu Y, Luo J, Ding P (2018) Inferring microrna targets based on restricted Boltzmann machines. IEEE J Biomed Health Inform 23(1):427–436
15. Jayashree R (2022) Enhanced classification using restricted boltzmann machine method in deep learning for covid-19. In: Understanding COVID-19: the role of computational intelligence. Springer, pp 425–446
16. Mohd Hasri NN, Wen NH, Howe CW, Mohamad MS, Deris S, Kasim S (2017) Improved support vector machine using multiple SVM-RFE for cancer classification. Int J Adv Sci Eng Inf Technol 7(4–2):1589–1594
17. Yuan X, Xie L, Abouelenien M (2018) A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data. Pattern Recognit 77:160–172
18. Gupta S, Kumar M (2021) Prostate cancer prognosis using multi-layer perceptron and class balancing techniques. In: 2021 13th international conference on contemporary computing (IC3-2021), pp 1–6
19. Ding H, Chen L, Dong L, Fu Z, Cui X (2022) Imbalanced data classification: A KNN and generative adversarial networks-based hybrid approach for intrusion detection. Future Gener Comput Syst 131:240–254
20. Qu X, Yang L, Guo K, Ma L, Sun M, Ke M, Li M (2021) A survey on the development of self-organizing maps for unsupervised intrusion detection. Mobile Netw Appl 26(2):808–829
21. Aldwairi T, Perera D, Novotny MA (2018) An evaluation of the performance of restricted Boltzmann machines as a model for anomaly network intrusion detection. Comput Netw 144:111–119

22. Gupta N, Jindal V, Bedi P (2021) LIO IDS: handling class imbalance using LSTM and improved one-vs-one technique in intrusion detection system. Comput Netw 192:1080–76
23. Pal A, Kumar M (2019) DLME: distributed log mining using ensemble learning for fault prediction. IEEE Syst J 13(4):3639–3650
24. Liu S, Jiang H, Wu Z, Li X (2022) Data synthesis using deep feature enhanced generative adversarial networks for rolling bearing imbalanced fault diagnosis. Mechan Syst Signal Process 163:108139
25. Peng Y, Wang Y, Shao Y (2022) A novel bearing imbalance fault-diagnosis method based on a wasserstein conditional generative adversarial network. Measurement 192:110924
26. Zhang W, Li X, Jia XD, Ma H, Luo Z, Li X (2020) Machinery fault diagnosis with imbalanced data using deep generative adversarial networks. Measurement 152:107377
27. Jang J, Kim CO (2022) Unstructured borderline self-organizing map: learning highly imbalanced, high-dimensional datasets for fault detection. Expert Syst Appl 188:116028
28. Kim JK, Lee JS, Han YS (2019) Fault detection prediction using a deep belief network-based multi-classifier in the semiconductor manufacturing process. Int J Softw Eng Knowl Eng 29:1125–1139
29. Peng P, Zhang W, Zhang Y, Wang H, Zhang H (2022) Non-revisiting genetic cost-sensitive sparse autoencoder for imbalanced fault diagnosis. Appl Soft Comput 114:108138
30. Zhao B, Zhang X, Li H, Yang Z (2020) Intelligent fault diagnosis of rolling bearings based on normalized CNN considering data imbalance and variable working conditions. Knowl Based Syst 199:105971
31. Zhu J, Jiang Q, Shen Y, Qian C, Xu F, Zhu Q (2022) Application of recurrent neural network to mechanical fault diagnosis: a review. J Mechan Sci Technol 36(2):1–16
32. Liu J, Zhang C, Jiang X (2022) Imbalanced fault diagnosis of rolling bearing using improved MsR-GAN and feature enhancement-driven CapsNet. Mechan Syst Signal Process 168
33. Dangut MD, Skaf Z, Jennions IK (2022) Handling imbalanced data for aircraft predictive maintenance using the BACHE algorithm. Appl Soft Comput 123:108924
34. De S, Prabu P (2022) A sampling-based stack framework for imbalanced learning in churn prediction. IEEE Access 10:68017–68028
35. Toor AA, Usman M (2022) Adaptive telecom churn prediction for concept-sensitive imbalance data streams. J Supercomput 78(3):3746–3774
36. Kimura T (2022) Customer churn prediction with hybrid resampling and ensemble learning. J Manag Inf Decis Sci 25(1)
37. Edwine N, Wang W, Song W, Ssebuggwawo D (2022) Detecting the risk of customer churn in telecom sector: a comparative study. Math Probl Eng 2022
38. Johnson JM, Khoshgoftaar TM (2019) Survey on deep learning with class imbalance. J Big Data 6(1):1–54
39. Moghar A, Hamiche M (2020) Stock market prediction using LSTM recurrent neural network. Procedia Comput Sci 170:1168–1173
40. Akşehir ZD, Kiliç E (2022) How to handle data imbalance and feature selection problems in CNN-based stock price forecasting. IEEE Access 10:31297–31305
41. Wang X, Zhang R, Zhang Z (2022) A novel hybrid sampling method esmote+ sslm for handling the problem of class imbalance with overlap in financial distress detection. Neural Process Lett, pp 1–25
42. Wu JM-T, Li Z, Srivastava G, Tasi MH, Lin JCW (2021) A graph-based convolutional neural network stock price prediction with leading indicators. Softw Pract Exp 51(3):628–644
43. Kawintiranon K, Singh L, Budak C (2022) Traditional and context-specific spam detection in low resource settings. Mach Learn 111(7):1–22
44. Wang G, Wang J, He K (2022) Majority-to-minority resampling for boosting-based classification under imbalanced data. Appl Intell 53(4):1–22
45. Lingam G, Yasaswini B, Jagadamba PVSL, Kolliboyana N (2022) An improved bot identification with imbalanced data using GG-XGBoost. In: 2022 2nd International conference on intelligent technologies (CONIT), pp 1–6
46. Hazarika BB, Gupta D (2022) Density weighted twin support vector machines for binary class imbalance learning. Neural Process Lett 54(2):1091–1130
47. Hossain T, Mauni HZ, Rab R (2022) Reducing the effect of imbalance in text classification using SVD and glove with ensemble and deep learning. Comput Inform 41(1):98–115
48. Rashid MRU, Mahbub M, Adnan MA (2022) Breaking the curse of class imbalance: bangla text classification. Trans Asian Low-Resour Lang Inf Process 21(5):1–21
49. Khurana A, Verma OP (2022) Optimal feature selection for imbalanced text classification. IEEE Trans Artif Intell
50. Wang Z, Wang H (2021) Global data distribution weighted synthetic oversampling technique for imbalanced learning. IEEE Access 9:44770–44783

51. Epasto A, Lattanzi S, Leme RP (2017) Ego-splitting framework: from non-overlapping to overlapping clusters. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pp 145–154
52. Wang S, Yao X (2009) Diversity analysis on imbalanced data sets by using ensemble models. In: 2009 IEEE symposium on computational intelligence and data mining, pp 324–331
53. Lu Y, Cheung Y-M, Tang YY (2016) Hybrid sampling with bagging for class imbalance learning. In: Pacific-Asia conference on knowledge discovery and data mining, pp 14–26
54. Wang S, Yao X (2009) Diversity analysis on imbalanced data sets by using ensemble models. In: 2009 IEEE symposium on computational intelligence and data mining, pp 324–331
55. Zhao Y, Liu S, Hu Z (2022) Focal learning on stranger for imbalanced image segmentation. IET Image Process 16(5):1305–1323
56. Ruwani K, Fernando M, Tsokos CP (2021) Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks. IEEE Trans Neural Netw Learn Syst
57. Jeong JJ, Tariq A, Adejumo T, Trivedi H, Gichoya JW, Banerjee I (2022) Systematic review of generative adversarial networks (GANs) for medical image classification and segmentation. J Digit Imag 35:1–16
58. Stoyanov D, Taylor Z, Carneiro G, Syeda-Mahmood T, Martel A, Maier-Hein L, Tavares JMRS, Bradley A, Papa JP, Belagiannis V et al (2018) Deep learning in medical image analysis and multimodal learning for clinical decision support. In: 4th International workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, Held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings, vol 11045. Springer
59. Akil M, Saouli R, Kachouri R et al (2020) Fully automatic brain tumor segmentation with deep learning-based selective attention using overlapping patches and multi-class weighted cross-entropy. Med Image Anal 63:101692
60. Nyo MT, Mebarek-Oudina F, Hlaing SS, Khan NA (2022) Otsu's thresholding technique for mri image brain tumor segmentation. Multimedia Tools Appl 81(30):43837–43849
61. Sampath V, Maurtua I, Aguilar Martín JJ, Gutierrez A (2021) A survey on generative adversarial networks for imbalance problems in computer vision tasks. J Big Data 8:1–59
62. Fendri E, Hammami M (2022) Imbalanced learning for robust moving object classification in video surveillance applications. In: Intelligent systems design and applications: 21st international conference on intelligent systems design and applications (ISDA 2021) held during december 13–15, 2021. Springer, vol 418, pp 199
63. Zhang Y, Lin M, Yang Y, Ding C (2022) A hybrid ensemble and evolutionary algorithm for imbalanced classification and its application on bioinformatics. Comput Biol Chem 98:107646
64. Dou L, Yang F, Xu L, Zou Q (2021) A comprehensive review of the imbalance classification of protein post-translational modifications. Brief Bioinform 22(5):bbab089
65. Thavappiragasam M, Kale V, Hernandez O, Sedova A (2021) Addressing load imbalance in bioinformatics and biomedical applications: efficient scheduling across multiple GPUs. In: 2021 IEEE international conference on bioinformatics and biomedicine (BIBM), pp 1992–1999
66. Chen J, Yang R, Zhang C, Zhang L, Zhang Q (2019) DeepGly: a deep learning framework with recurrent and convolutional neural networks to identify protein glycation sites from imbalanced data. IEEE Access 7:142368–142378
67. Greene CS, Himmelstein DS, Kiralis J, Moore JH (2010) The informative extremes: using both nearest and farthest individuals can improve relief algorithms in the domain of human genetics. In: European conference on evolutionary computation, machine learning and data mining in bioinformatics, pp 182–193
68. Greene CS, Penrod NM, Kiralis J, Moore JH (2009) Spatially uniform relieff (SURF) for computationally-efficient filtering of gene-gene interactions. BioData Min 2(1):1–9
69. Djenouri Y, Belhadi A, Srivastava G, Lin JCW (2021) Secure collaborative augmented reality framework for biomedical informatics. IEEE J Biomed Health Inform 26(6):2417–2424
70. Chen L, Fang B, Shang Z, Tang Y (2018) Tackling class overlap and imbalance problems in software defect prediction. Softw Qual J 26(1):97–125
71. Goyal S (2022) Handling class-imbalance with KNN (neighbourhood) under-sampling for software defect prediction. Artif Intell Rev 55(3):2023–2064
72. Manchala P, Bisi M (2022) Diversity based imbalance learning approach for software fault prediction using machine learning models. Appl Soft Comput 124:109069
73. Yin J, Tang MJ, Cao J, Wang H, You M, Lin Y (2022) Vulnerability exploitation time prediction: an integrated framework for dynamic imbalanced learning. World Wide Web 25(1):401–423
74. Lu S, Gao Z, Xu Q, Jiang C, Zhang A, Wang X (2022) Class-imbalance privacy-preserving federated learning for decentralized fault diagnosis with biometric authentication. IEEE Trans Ind Inform

75. Sun M, Yang R, Liu M (2022) Privacy-preserving minority oversampling protocols with fully homomorphic encryption. Secur Commun Netw 2022
76. Singh K, Mahajan A, Mansotra V (2022) Deep learning approach based on ADASYN for detection of web attacks in the CICIDS2017 dataset. In: Rising threats in expert applications and solutions. Springer, pp 53–62
77. Le TTH, Oktian YE, Kim H (2022) Xgboost for imbalanced multiclass classification-based industrial internet of things intrusion detection systems. Sustainability 14(14):8707
78. Zhang S, Yin J, Li Z, Yang R, Du M, Li R (2022) Node-imbalance learning on heterogeneous graph for pirated video website detection. In: 2022 IEEE 25th international conference on computer supported cooperative work in design (CSCWD). IEEE, pp 834–840
79. Santos MS, Abreu PH, Japkowicz N, Fernández A, Soares C, Wilk S, Santos J (2022) On the joint-effect of class imbalance and overlap: a critical review. Artif Intell Rev 55(8):1–69
80. Santos MS, Abreu PH, Japkowicz N, Fernández A, Santos J (2022) A unifying view of class overlap and imbalance: key concepts, multi-view panorama, and open avenues for research. Inf Fusion
81. Branco P, Torgo L, Ribeiro RP (2016) A survey of predictive modeling on imbalanced domains. ACM Comput Surv (CSUR) 49(2):1–50
82. Rout N, Mishra D, Mallick MK (2018) Handling imbalanced data: a survey. In: International proceedings on advances in soft computing, intelligent systems and applications. Springer, pp 431–443
83. Kaur H, Pannu HS, Malhi AK (2019) A systematic review on imbalanced data challenges in machine learning: applications and solutions. ACM Comput Surv (CSUR) 52(4):1–36
84. Xiong H, Wu J, Liu L (2010) Classification with class overlapping: a systematic study. In: 2010 International conference on E-business intelligence, pp 491–497
85. Liu X, Fu L, Lin JCW, Liu S (2022) SRAS-net: low-resolution chromosome image classification based on deep learning. IET Syst Biol 16(3–4):85–97
86. Tian C, Zhang X, Lin JCW, Zuo W, Zhang Y, Lin CW (2022) Generative adversarial networks for image super-resolution: a survey. arXiv:2204.13620
87. Mezair T, Djenouri Y, Belhadi A, Srivastava G, Lin JCW (2022) A sustainable deep learning framework for fault detection in 6G industry 4.0 heterogeneous data environments. Comput Commun 187:164–171
88. Akondi VS, Menon V, Baudry J, Whittle J (2022) Novel big data-driven machine learning models for drug discovery application. Molecules 27(3):594
89. Khattak A, Bukhsh R, Aslam S, Yafoz A, Alghushairy O, Alsini R (2022) A hybrid deep learning-based model for detection of electricity losses using big data in power systems. Sustainability 14(20):13627
90. Hewamalage H, Bergmeir C, Bandara K (2021) Recurrent neural networks for time series forecasting: current status and future directions. Int J Forecast 37:388–427
91. Das S, Datta S, Chaudhuri BB (2018) Handling data irregularities in classification: foundations, trends, and future challenges. Pattern Recognit 81:674–693
92. Napierała K, Stefanowski J, Wilk S (2010) Learning from imbalanced data in presence of noisy and borderline examples. In: International conference on rough sets and current trends in computing. Springer, pp 158–167
93. López V, Fernández A, García S, Palade V, Herrera F (2013) An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inf Sci 250:113–141
94. Stefanowski J (2016) Dealing with data difficulty factors while learning from imbalanced data. In: Challenges in computational statistics and data mining. Springer, pp 333–363
95. Wojciechowski S, Wilk S (2017) Difficulty factors and preprocessing in imbalanced data sets: an experimental study on artificial data. Found Comput Decis Sci 42(2):149–176
96. García V, Mollineda RA, Sánchez JS (2008) On the k-NN performance in a challenging scenario of imbalance and overlapping. Pattern Anal Appl 11(3):269–280
97. Lee HK, Kim SB (2018) An overlap-sensitive margin classifier for imbalanced and overlapping data. Expert Syst Appl 98:72–83
98. Das B, Krishnan NC, Cook DJ (2014) Handling imbalanced and overlapping classes in smart environments prompting dataset. In: Data Min Serv. Springer, pp 199–219
99. Pascual-Triana JD, Charte D, Arroyo MA, Fernández A, Herrera F (2021) Revisiting data complexity metrics based on morphology for overlap and imbalance: snapshot, new overlap number of balls metrics and singular problems prospect. Knowl Inf Syst 63:1–29
100. Vuttipittayamongkol P, Elyan E (2020) Improved overlap-based undersampling for imbalanced dataset classification with application to Epilepsy and Parkinson's disease. Int J Neural Syst 30(08):2050043
101. Dkhar RA, Nath K, Roy S, Bhattacharyya DK, Nandi S (2016) Evaluating the effectiveness of soft k-means in detecting overlapping clusters. In: Proceedings of the 2nd international conference on information and communication technology for competitive strategies, pp 1–6

102. Tao X, Chen W, Zhang X, Guo W, Qi L, Fan Z (2021) SVDD boundary and DPC clustering technique-based oversampling approach for handling imbalanced and overlapped data. Knowl Based Syst 234:107588
103. Xiong H, Li M, Jiang T, Zhao S (2013) Classification algorithm based on nb for class overlapping problem. Appl Math 7(2L):409–415
104. Tung NT, Dieu VH, Than K, Linh NV (2018) Reducing class overlapping in supervised dimension reduction. In: Proceedings of the 9th international symposium on information and communication technology, pp 8–15
105. Fernandes ERQ, De Carvalho AC (2019) Evolutionary inversion of class distribution in overlapping areas for multi-class imbalanced learning. Inf Sci 494:141–154
106. Li Z, Huang M, Liu G, Jiang C (2021)A hybrid method with dynamic weighted entropy for handling the problem of class imbalance with overlap in credit card fraud detection. Expert Syst Appl 175:114750
107. Wong ML, Seng K, Wong PK (2020) Cost-sensitive ensemble of stacked denoising autoencoders for class imbalance problems in business domain. Expert Syst Appl 141:112918
108. Rogić S, Kašćelan L, Bach MP (2022) Customer response model in direct marketing: solving the problem of unbalanced dataset with a balanced support vector machine. J Theor Appl Electron Commer Res 17(3):1003–1018
109. Zhu B, Pan X, Vanden Broucke S, Xiao J (2022) A GAN-based hybrid sampling method for imbalanced customer classification. Inf Sci 609:1397–1411
110. Ntomaris AV, Marneris IG, Biskas PN, Bakirtzis AG (2022) Optimal participation of RES aggregators in electricity markets under main imbalance pricing schemes: price taker and price maker approach. Electr Power Syst Res 206:107786
111. Lee D, Kim K (2022) Business transaction recommendation for discovering potential business partners using deep learning. Expert Syst Appl 201:117222
112. Garcia J (2022) Bankruptcy prediction using synthetic sampling. Mach Learn Appl 9:100343
113. Rodić LD, Perković T, Škiljo M, Šolić P (2022) Privacy leakage of lorawan smart parking occupancy sensors. Future Gener Comput Syst
114. Vuttipittayamongkol P, Elyan E (2020) Overlap-based undersampling method for classification of imbalanced medical datasets. In: Maglogiannis I, Iliadis L, Pimenidis E (eds) Artificial intelligence applications and innovations. Springer, Cham, pp 358–369
115. Zhang R, Zhang Z, Wang D (2021) RFCL: a new under-sampling method of reducing the degree of imbalance and overlap. Pattern Anal Appl 24(2):641–654
116. Devi D, Biswas SK, Purkayastha B (2019) Learning in presence of class imbalance and class overlapping by using one-class SVM and undersampling technique. Connect Sci 31(2):105–142
117. Vuttipittayamongkol P, Elyan E, Petrovski A, Jayne C (2018) Overlap-based undersampling for improving imbalanced data classification. In: International conference on intelligent data engineering and automated learning. Springer, pp 689–697
118. Ibrahim MH (2021) ODBOT: outlier detection-based oversampling technique for imbalanced datasets learning. Neural Comput Appl 33:15781–15806
119. Tao X, Zheng Y, Chen W, Zhang X, Qi L, Fan Z, Huang S (2022) SVDD-based weighted oversampling technique for imbalanced and overlapped dataset learning. Inf Sci 588:13–51
120. Zhu Y, Yan Y, Zhang Y, Zhang Y (2020) EHSO: evolutionary hybrid sampling in overlapping scenarios for imbalanced learning. Neurocomputing 417:333–346
121. Maldonado S, Vairetti C, Fernandez A, Herrera F (2022) FW-SMOTE: a feature-weighted oversampling approach for imbalanced classification. Pattern Recognit 124:108511
122. Tax DMJ, Duin RPW (2004) Support vector data description. Mach Learn 54(1):45–66
123. Mayabadi S, Saadatfar H (2022) Two density-based sampling approaches for imbalanced and overlapping data. Knowl Based Syst 241:108217
124. Zian S, Kareem SA, Varathan KD (2021) An empirical evaluation of stacked ensembles with different meta-learners in imbalanced classification. IEEE Access
125. Sumana BV, Punithavalli M (2020) Optimising prediction in overlapping and non-overlapping regions. Int J Nat Comput Res (IJNCR) 9(1):45–63
126. Gupta S, Gupta A (2018) Handling class overlapping to detect noisy instances in classification. Knowl Eng Rev 33
127. Chujai P, Chomboon K, Chaiyakhan K, Kerdprasop K, Kerdprasop N (2017) A cluster based classification of imbalanced data with overlapping regions between classes. Proceedings of the international multiconference of engineers and computer scientists 1:353–358
128. Liu C, Ren Y, Liang M, Gu Z, Wang J, Pan L, Wang Z (2020) Detecting overlapping data in system logs based on ensemble learning method. Wireless Commun Mobile Comput 2020:1–8

129. De Miguel L, Gómez D, Rodríguez JT, Montero J, Bustince H, Dimuro GP, Sanz JA (2019) General overlap functions. Fuzzy Sets Syst 372:81–96
130. Elkan C (2001) The foundations of cost-sensitive learning. International joint conference on artificial intelligence, vol 17. Lawrence Erlbaum Associates Ltd, Mahwah, pp 973–978
131. Xia Y, Liu C, Liu N (2017) Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. Electron Commer Res Appl 24:30–49
132. Yang S, Korayem M, AlJadda K, Grainger T, Natarajan S (2017) Combining content-based and collaborative filtering for job recommendation system: a cost-sensitive statistical relational learning approach. Knowl Based Syst 136:37–45
133. Yuan BW, Luo XG, Zhang ZL, Yu Y, Huo HW, Johannes T, Zou XD (2021) A novel density-based adaptive k nearest neighbor method for dealing with overlapping problem in imbalanced datasets. Neural Comput Appl 33(9):4457–4481
134. Rubbo M, Silv LA (2021) Filtering-based instance selection method for overlapping problem in imbalanced datasets. J 4(3):308–327
135. Zhang N, Karimoune W, Thompson L, Dang H (2017) A between-class overlapping coherence-based algorithm in KNN classification. In: 2017 IEEE international conference on systems, man, and cybernetics (SMC). IEEE, pp 572–577
136. Gu Y, Cheng L (2017) Classification of class overlapping datasets by kernel-MTS method. Int J Innovat Comput Inf Control 13(5):1759–1767
137. Afridi MK, Azam N, Yao J (2020) Variance based three-way clustering approaches for handling overlapping clustering. Int J Approx Reason 118:47–63
138. Li H, Zhang L, Zhou X, Huang B (2017) Cost-sensitive sequential three-way decision modeling using a deep neural network. Int J Approx Reason 85:68–78
139. Lee HK, Kim SB (2018) An overlap-sensitive margin classifier for imbalanced and overlapping data. Expert Syst Appl 98:72–83
140. Mienye ID, Sun Y (2021) Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. Inf Med Unlocked 25:100690
141. Lin X, Li C, Zhang Y, Su B, Fan M, Wei H (2018) Selecting feature subsets based on svm-rfe and the overlapping ratio with applications in bioinformatics. Molecules 23(1):52
142. Akhter S, Sharmin S, Ahmed S, Sajib AA, Shoyaib M (2021) mRelief: a reward penalty based feature subset selection considering data overlapping problem. In: International conference on computational science. Springer, pp 278–292
143. Omar B, Rustam F, Mehmood A, Choi GS (2021) Minimizing the overlapping degree to improve class-imbalanced learning under sparse feature selection: application to fraud detection. IEEE Access 9:28101–28110
144. Alshomrani S, Bawakid A, Shim Seong-O, Fernández A, Herrera F (2015) A proposal for evolutionary fuzzy systems using feature weighting: dealing with overlapping in imbalanced datasets. Knowl Based Syst 73:1–17
145. Zhang Y, Cheng S, Shi Y, Gong DW, Zhao X (2019) Cost-sensitive feature selection using two-archive multi-objective artificial bee colony algorithm. Expert Syst Appl 137:46–58
146. Sáez JA, Galar M, Krawczyk B (2019) Addressing the overlapping data problem in classification using the one-vs-one decomposition strategy. IEEE Access 7:83396–83411
147. Shahee SA, Ananthakumar U (2021) An overlap sensitive neural network for class imbalanced data. Data Min Knowl Discov 35(4):1–34
148. Yuan BW, Zhang ZL, Luo XG, Yu Y, Zou XH, Zou XD (2021) OIS-RF: a novel overlap and imbalance sensitive random forest. Eng Appl Artif Intell 104:104355
149. Nwe MM, Lynn KT (2019) kNN-based overlapping samples filter approach for classification of imbalanced data. In: International conference on software engineering research, management and applications. Springer, pp 55–73
150. Yan Y, Jiang Y, Zheng Z, Yu C, Zhang Y, Zhang Y (2022) LDAS: local density-based adaptive sampling for imbalanced data classification. Expert Syst Appl 191:116213
151. Roy A, Cruz RM, Sabourin R, Cavalcanti GD (2018) A study on combining dynamic selection and data preprocessing for imbalance learning. Neurocomputing 286:179–192
152. Pouyanfar S, Sadiq S, Yan Y, Tian H, Tao Y, Reyes MP, Shyu ML, Chen SC, Iyengar SS (2018) A survey on deep learning: algorithms, techniques, and applications. ACM Comput Surv (CSUR) 51:1–36
153. Tong K, Wu Y (2022) Deep learning-based detection from the perspective of small or tiny objects: a survey. Image Vis Comput 123:104471
154. Liu Z, Tong L, Jiang Z, Chen L, Zhou F, Zhang Q, Zhang X, Jin Y, Zhou H (2020) Deep learning based brain tumor segmentation: a survey. Preprint at https://arxiv.org/abs/2007.09479

155. Wong LJ, Headley WC, Michaels AJ (2019) Specific emitter identification using convolutional neural network-based IQ imbalance estimators. IEEE Access 7:33544–33555
156. Chen Z, Duan J, Kang L, Qiu G (2021) Class-imbalanced deep learning via a class-balanced ensemble. IEEE Trans Neural Netw Learn Syst
157. Yan Y, Chen M, Shyu ML, Chen SC (2015) Deep learning for imbalanced multimedia data classification. In: 2015 IEEE international symposium on multimedia (ISM). IEEE, pp 483–488
158. Böhm A, Ücker A, Jäger T, Ronneberger O, Falk T (2018) ISOO_DL: Instance segmentation of overlapping biological objects using deep learning. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). IEEE, pp 1225–1229
159. Banerjee I, Ling Y, Chen MC, Hasan SA, Langlotz CP, Moradzadeh N, Chapman B, Amrhein T, Mong D, Rubin DL (2019) Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. Artif Intell Med 97:79–88
160. Gao L, Lu P, Ren Y (2021) A deep learning approach for imbalanced crash data in predicting highway-rail grade crossings accidents. Reliab Eng Syst Saf 216:108019
161. Rai HM, Chatterjee K (2022) Hybrid CNN LSTM deep learning model and ensemble technique for automatic detection of myocardial infarction using big ECG data. Appl Intell 52(5):5366–5384
162. Gao J, Zhang H, Lu P, Wang Z (2019) An effective LSTM recurrent network to detect arrhythmia on imbalanced ecg dataset. J Healthc Eng
163. Tran D, Mac H, Tong V, Tran HA, Nguyen LG (2018) A LSTM based framework for handling multiclass imbalance in DGA botnet detection. Neurocomputing 275:2401–2413
164. Dong Q, Gong S, Zhu X (2018) Imbalanced deep learning by minority class incremental rectification. IEEE Trans Pattern Anal Mach Intell 41(6):1367–1381
165. Lin E, Chen Q, Qi X (2020) Deep reinforcement learning for imbalanced classification. Appl Intell 50(8):2488–2502
166. Wang S, Liu W, Wu J, Cao L, Meng Q, Kennedy PJ (2016) Training deep neural networks on imbalanced data sets. In: 2016 International joint conference on neural networks (IJCNN). IEEE, pp 4368–4374
167. Zhang C, Tan KC, Li H, Hong GS (2018) A cost-sensitive deep belief network for imbalanced classification. IEEE Trans Neural Netw Learn Syst 30(1):109–122
168. Dablain D, Krawczyk B, Chawla NV (2022) DeepSMOTE: fusing deep learning and smote for imbalanced data. IEEE Trans Neural Netw Learn Syst
169. Andrei V, Cucu H, Burileanu C (2019) Overlapped speech detection and competing speaker counting–humans versus deep learning. IEEE J Sel Topics Signal Process 13(4):850–862
170. Alia A, Maree M, Chraibi M (2022) A hybrid deep learning and visualization framework for pushing behavior detection in pedestrian dynamics. Sensors 22(11):4040
171. Wang X, Jing L, Lyu Y, Guo M, Wang J, Liu H, Yu J, Zeng T (2022) Deep generative mixture model for robust imbalance classification. IEEE Trans Pattern Anal Mach Intell
172. Yue X, Li H, Fujikawa Y, Meng L (2022) Dynamic dataset augmentation for deep learning-based oracle bone inscriptions recognition. J Comput Cult Herit (JOCCH)
173. Liu T, Bao J, Wang J, Wang J (2021) Deep learning for industrial image: challenges, methods for enriching the sample space and restricting the hypothesis space, and possible issue. Int J Comput Integr Manuf 35:1–30
174. ArunKumar KE, Kalaga DV, Kumar CMS, Kawaji M, Brenza TM (2021) Forecasting of covid-19 using deep layer recurrent neural networks (RNNs) with gated recurrent units (GRUs) and long short term memory (LSTM) cells. Chaos, Solitons Fractals 146:110861
175. Zhang Q, Wang W, Zhu SC (2018) Examining cnn representations with respect to dataset bias. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
176. Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv:1702.08608
177. Ibrahim M, Louie M, Modarres C, Paisley J (2019) Global explanations of neural networks: mapping the landscape of predictions. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society, pp 279–287
178. Wu JMT, Li Z, Herencsar N, Vo B, Lin JCW (2021) A graph-based CNN-LSTM stock price prediction algorithm with leading indicators. Multimedia Syst 29:1–20
179. Sherstinsky A (2020) Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Phys D Nonlin Phenom 404:132306
180. Chen MY, Chiang HS, Huang WK (2022) Efficient generative adversarial networks for imbalanced traffic collision datasets. IEEE Trans Intell Transp Syst
181. Lee HK, Lee J, Kim SB (2022) Boundary-focused generative adversarial networks for imbalanced and multimodal time series. IEEE Trans Knowl Data Eng

182. Li W, Chen J, Cao J, Ma C, Wang J, Cui X, Chen P (2022) EID-GAN: generative adversarial nets for extremely imbalanced data augmentation. IEEE Trans Ind Inform
183. Gao S, Dai Y, Li Y, Liu K, Chen K, Liu Y (2022) Multiview wasserstein generative adversarial network for imbalanced pearl classification. Meas Sci Technol 33(8):085406
184. Suh S, Lee H, Lukowicz P, Lee YO (2021) CEGAN: classification enhancement generative adversarial networks for unraveling data imbalance problems. Neural Netw 133:69–86
185. De Oliveira Nogueira T, Palacio GBA, Braga FD, Maia PPN, De Moura EP, De Andrade CF, Rocha PAC (2022) Imbalance classification in a scaled-down wind turbine using radial basis function kernel and support vector machines. Energy 238:122064
186. Satapathy SK, Mishra S, Mallick PK, Chae GS (2021) ADASYN and ABC-optimized RBF convergence network for classification of electroencephalograph signal. Pers Ubiquitous Comput 27:1–17
187. Zhang D, Zhang N, Ye N, Fang J, Han X (2020) Hybrid learning algorithm of radial basis function networks for reliability analysis. IEEE Trans Reliab 70(3):887–900
188. Kamaruddin SK, Ravi V (2019) A parallel and distributed radial basis function network for big data analytics. In: TENCON 2019-2019 IEEE Region 10 Conference (TENCON). IEEE, pp 395–399
189. Akter S, Das D, Haque RU, Tonmoy MIQ, Hasan MR, Mahjabeen S, Ahmed M (2022) AD-covNet: an exploratory analysis using a hybrid deep learning model to handle data imbalance, predict fatality, and risk factors in Alzheimer's patients with covid-19. Comput Biol Med 146:105657
190. Ram PK, Kuila P (2022) GAAE: a novel genetic algorithm based on autoencoder with ensemble classifiers for imbalanced healthcare data. J Supercomput 79:1–32
191. Hassib EM, El-Desouky AI, Labib LM, El-Kenawy ESM (2020) WOA+BRNN: an imbalanced big data classification framework using whale optimization and deep neural network. Soft Comput 24(8):5573–5592
192. Dumas J, Boukas I, De Villena MM, Mathieu S, Cornélusse B (2019) Probabilistic forecasting of imbalance prices in the Belgian context. In: 2019 16th International conference on the European energy market (EEM). IEEE, pp 1–7
193. Ghanem WA, Jantan A (2018) A cognitively inspired hybridization of artificial bee colony and dragonfly algorithms for training multi-layer perceptrons. Cogn Comput 10(6):1096–1134
194. Zhu G, Wu X, Ge J, Liu F, Zhao W, Wu C (2020) Influence of mining activities on groundwater hydrochemistry and heavy metal migration using a self-organizing map (SOM). J Clean Prod 257:120664
195. Hameed AA, Karlik B, Salman MS, Eleyan G (2019) Robust adaptive learning approach to self-organizing maps. Knowl Based Syst 171:25–36
196. Huysmans D, Smets E, De Raedt W, Van Hoof C, Bogaerts K, Van Diest I, Helic D (2018) Unsupervised learning for mental stress detection-exploration of self-organizing maps. Proceedings of the 11th international joint conference on biomedical engineering systems and technologies, vol 4, pp 26–35
197. Xie H, Wu L, Xie W, Lin Q, Liu M, Lin Y (2021) Improving ECMWF short-term intensive rainfall forecasts using generative adversarial nets and deep belief networks. Atmos Res 249:105281
198. Vinayakumar R, Alazab M, Srinivasan S, Pham QV, Padannayil SK, Simran K (2020) A visualized botnet detection system based deep learning for the internet of things networks of smart cities. IEEE Trans Ind Appl 56:4436–4456
199. Leonelli FE, Agliari E, Albanese L, Barra A (2021) On the effective initialisation for restricted Boltzmann machines via duality with Hopfield model. Neural Netw 143:314–326
200. Savitha R, Ambikapathi A, Rajaraman K (2020) Online RBM: growing restricted boltzmann machine on the fly for unsupervised representation. Appl Soft Comput 92:106278
201. Huang K, Wang X (2022) ADA-INCVAE: improved data generation using variational autoencoder for imbalanced classification. Appl Intell 52(3):2838–2853
202. Chen J, Wu Z, Zhang J (2019) Driving safety risk prediction using cost-sensitive with nonnegativity-constrained autoencoders based on imbalanced naturalistic driving data. IEEE Trans Intell Transp Syst 20(12):4450–4465
203. Alhassan Z, Budgen D, Alshammari R, Daghstani T, McGough AS, Al Moubayed N (2018) Stacked denoising autoencoders for mortality risk prediction using imbalanced clinical data. In: 2018 17th IEEE international conference on machine learning and applications (ICMLA). IEEE, pp 541–546
204. Johnson JM, Khoshgoftaar TM (2020) The effects of data sampling with deep learning and highly imbalanced big data. Inf Syst Front 22(5):1113–1131
205. Yan M, Li N (2022) Borderline-margin loss based deep metric learning framework for imbalanced data. Appl Intell 53:1–18
206. Lin WC, Tsai CF, Hu YH, Jhang JS (2017) Clustering-based undersampling in class-imbalanced data. Inf Sci 409:17–26
207. Vannucci M, Colla V (2018) Self–organizing–maps based undersampling for the classification of unbalanced datasets. In: 2018 International joint conference on neural networks (IJCNN). IEEE, pp 1–6

208. Tsai CF, Lin WC, Hu YH, Yao GT (2019) Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. Inf Sci 477:47–54
209. More A (2016) Survey of resampling techniques for improving classification performance in unbalanced datasets. arXiv:1608.06048
210. Yang Z, Gao D (2013) Classification for imbalanced and overlapping classes using outlier detection and sampling techniques. Appl Math Inf Sci 7(1):375–381
211. Douzas G, Bacao F, Last F (2018) Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. Inf Sci 465:1–20
212. Barua S, Islam MM, Yao X, Murase K (2012) MWMOTE-majority weighted minority oversampling technique for imbalanced data set learning. IEEE Trans Knowl Data Eng 26(2):405–425
213. He H, Bai Y et al (2008) ADASYN: adaptive synthetic sampling for imbalanced data. In: IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), vol 69. https://doi.org/10.1109/ijcnn
214. Ren R, Yang Y, Sun L (2020) Oversampling technique based on fuzzy representativeness difference for classifying imbalanced data. Appl Intell 50(8):2465–2487
215. Elyan E, Moreno-Garcia CF, Jayne C (2021) CDSMOTE: class decomposition and synthetic minority class oversampling technique for imbalanced-data classification. Neural Comput Appl 33(7):2839–2851
216. Liu G, Yang Y, Li B (2018) Fuzzy rule-based oversampling technique for imbalanced and incomplete data learning. Knowl Based Syst 158:154–174
217. Koziarski M, Krawczyk B, Wozniak M (2019) Radial-based oversampling for noisy imbalanced data classification. Neurocomputing 343:19–33
218. Yan Y, Liu R, Ding Z, Du X, Chen J, Zhang Y (2019) A parameter-free cleaning method for SMOTE in imbalanced classification. IEEE Access 7:23537–23548
219. Patel H, Thakur GS (2016) A hybrid weighted nearest neighbor approach to mine imbalanced data. In: Proceedings of the international conference on data science (ICDATA), The steering committee of the world congress in computer, science, Computer, pp 106
220. Tang B, He H (2015) ENN: extended nearest neighbor method for pattern recognition [research frontier]. IEEE Comput Intell Mag 10(3):52–60
221. Wang P, Yao Y (2018) CE3: a three-way clustering method based on mathematical morphology. Knowl Based Syst 155:54–65
222. Masson MH, Denoeux T (2009) RECM: relational evidential c-means algorithm. Pattern Recognit Lett 30(11):1015–1026
223. Batista GE, Prati RC, Monard MC (2004) A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor Newsl 6(1):20–29
224. Fan Q, Wang Z, Li D, Gao D, Zha H (2017) Entropy-based fuzzy support vector machine for imbalanced datasets. Knowl Based Syst 115:87–99
225. Zhu C, Wang Z (2017) Entropy-based matrix learning machine for imbalanced data sets. Pattern Recognit Lett 88:72–80
226. Wang BX, Japkowicz N (2010) Boosting support vector machines for imbalanced data sets. Knowl Inf Syst 25(1):1–20
227. Ju H, Li H, Yang X, Zhou X, Huang B (2017) Cost-sensitive rough set: a multi-granulation approach. Knowl Based Syst 123:137–153
228. Ju H, Yang X, Yu H, Li T, Yu DJ, Yang J (2016) Cost-sensitive rough set approach. Inf Sci 355:282–298
229. Cabitza F, Ciucci D, Locoro A (2017) Exploiting collective knowledge with three-way decision theory: cases from the questionnaire-based research. Int J Approx Reason 83:356–370
230. Maulidevi NU, Surendro K (2021) SMOTE-LOF for noise identification in imbalanced data classification. J King Saud Univ Comput Inf Sci
231. Armano G, Tamponi E (2018) Building forests of local trees. Pattern Recognit 76:380–390
232. Galar M, Fernández A, Barrenechea E, Herrera F (2013) EUSboost: enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. Pattern Recognit 46(12):3460–3471
233. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F (2011) A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans Syst, Man, Cybern, Part C (Appl Rev) 42(4):463–484
234. Sesmero MP, Ledezma AI, Sanchis A (2015) Generating ensembles of heterogeneous classifiers using stacked generalization. Wiley Interdiscip Rev Data Min Knowl Discov 5(1):21–34
235. Kim S, Zhang H, Wu R, Gong L (2011) Dealing with noise in defect prediction. In: 2011 33rd International conference on software engineering (ICSE). IEEE, pp 481–490
236. Tang W, Khoshgoftaar TM (2004) Noise identification with the k-means algorithm. In: 16th IEEE international conference on tools with artificial intelligence. IEEE, pp 373–378

237. Sundqvist T, Bhuyan MH, Forsman J, Elmroth E (2020) Boosted ensemble learning for anomaly detection in 5G RAN. In: IFIP international conference on artificial intelligence applications and innovations. Springer, pp 15–30
238. Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A (2009) RUSBoost: a hybrid approach to alleviating class imbalance. IEEE Trans Syst Man Cybern Part A Syst Hum 40(1):185–197
239. Tosin MC, Majolo M, Chedid R, Cene VH, Balbinot A (2017) sEMG feature selection and classification using SVM-RFE. In: 2017 39th Annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, pp 390–393
240. Alcala-Fdez J, Alcala R, Herrera F (2011) A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning. IEEE Trans Fuzzy Syst 19(5):857–872
241. Akhter S, Sharmin S, Ahmed S, Sajib AA, Shoyaib M (2021) mRelief: a reward penalty based feature subset selection considering data overlapping problem. In: International conference on computational science. Springer, pp 278–292
242. Min F, Hu Q, Zhu W (2014) Feature selection with test cost constraint. Int J Approx Reason 55(1):167–179
243. Zhao H, Wang P, Hu Q (2016) Cost-sensitive feature selection based on adaptive neighborhood granularity with multi-level confidence. Inf Sci 366:134–149
244. Emekter R, Tu Y, Jirasakuldech B, Lu M (2015) Evaluating credit risk and loan performance in online peer-to-peer (P2P) lending. Appl Econ 47(1):54–70
245. Vorraboot P, Rasmequan S, Chinnasarn K, Lursinsap C (2015) Improving classification rate constrained to imbalanced data between overlapped and non-overlapped regions by hybrid algorithms. Neurocomputing 152:429–443

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