International Journal of Advanced Technology and Engineering Exploration, Vol 10(107)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2023.10101138

Research Article

Performance analysis of samplers and calibrators with various classifiers for asymmetric hydrological data

C. Kaleeswari[1*], K. Kuppusamy[2] and A. Senthilrajan[3]
Research Scholar, Department of Computational Logistics, Alagappa University, Karaikudi, Tamilnadu, India[1]
Professor, Department of Computational Logistics, Alagappa University, Karaikudi, Tamilnadu, India[2]
Professor and Head, Department of Computational Logistics, Alagappa University, Karaikudi, Tamilnadu, India[3]

Received: 15-February-2023; Revised: 28-October-2023; Accepted: 29-October-2023

Abstract
Asymmetric data classification presents a significant challenge in machine learning (ML). While ML algorithms are known for their ability to classify symmetric data effectively, addressing data asymmetry remains an on-going concern in classification tasks. This research paper aims to select an appropriate method for classifying and predicting asymmetric data, focusing on label and probability predictions. To achieve this, various ML classifiers, calibration techniques, and sampling methods are systematically analyzed. The classifiers under consideration include logistic regression (LR), k-nearest neighbour (KNN), gaussian naive Bayes (GNB), random forest (RF), decision tree (DT), and support vector classifier (SVC). Calibration techniques explored encompass isotonic regression (IR) and platt scaling (PS), while sampling techniques comprise synthetic minority oversampling technique (SMOTE), T-link (Tomek), adaptive synthetic sampling (AdaSyn), integration of SMOTE and edited nearest neighbour (SMOTEENN), and integration of SMOTE and T-link (SMOTETomek). Simulation results for label prediction consistently favour the SMOTEENN approach, with the RF classifier combined with SMOTEENN providing outstanding performance, boasting a balanced random accuracy (BRA) of 98.07%, sensitivity of 98.02%, specificity of 99.01%, an area under the curve (AUC) of 0.98, and a geometric mean (G-mean) of 98.50%. In terms of probability prediction, IR calibration consistently excels. Specifically, the GNB classifier combined with IR produces the best performance, yielding a low brier score (BS), expected calibration error (ECE), and maximum calibration error (MCE). Furthermore, it achieves perfect calibration as demonstrated by the reliability curve. In light of these findings, this study recommends the utilization of SMOTEENN for data resampling and IR calibration for probability prediction as superior methods to address data asymmetry. The comparative analysis presented in this research offers valuable insights for selecting appropriate techniques in the context of asymmetric data classification.

Keywords
Machine learning, Calibration, Asymmetric data, Classification, Probability, Prediction.

1.Introduction
Water plays a crucial role in supporting the survival of all species on the globe. Access to clean, pristine water is essential for both humans and the environment [1, 2]. Identifying potential sources of pollution is of great significance in predicting water contamination, as human activities have placed the quality of water at risk [3]. Over time, there has been a growing variety of water quality analyses. This has led to an increasing reliance on computational algorithms for predicting water quality and contamination.

Often, these computational models and their applications need to address the challenges posed by unbalanced sensor data [4]. Dealing with such sensor data is known to involve several difficulties, including issues related to clutter and extreme asymmetry [5]. Classifying asymmetric data presents a challenging task that has attracted interest across various scientific disciplines [6]. Models constructed with bias from asymmetric datasets yield unreliable predictions and unsatisfactory classification results [7].

Prediction can be approached through various methods, each with its unique effectiveness [8]. A literature review reveals that most prediction studies have research gaps in the pursuit of more accurate

*Author for correspondence

and applicable models for forecasting asymmetric data [9, 10]. Asymmetric data used for prediction often hampers classifier performance and leads to various issues, including misclassification, under fitting, and exaggerated results. This is due to the presence of anomalies, missing values, and infinite values [11–16]. Consequently, for asymmetric data classification tasks, samples are typically categorized into two groups: the majority and minority classes [17]. In general, the minority class samples and interests hold greater significance and importance than those of the majority class. However, the majority class encompasses a larger number of samples than the minority class, and in some cases, this scenario can be quite challenging. Therefore, effectively addressing these challenges has emerged as a critical and essential subject within the realms of machine learning (ML) and deep learning (DL).

ML is well-suited for developing algorithms that can adapt and enhance their forecasting abilities by utilizing symmetric datasets [18, 19]. The performance of any ML method that excels depends on the data and the specific application. Asymmetric data poses a significant challenge in the field of ML, as it can negatively impact the performance of classification models [20]. Irrespective of the data or application, there are instances where models can yield favorable results with minimal loss. In such cases, it is crucial to identify and address over fitting issues [21]. Before applying ML approaches to develop classifiers, it is necessary to reduce the dimensionality of these asymmetric datasets [22]. During the preprocessing phase, various techniques, such as anomalous data mining, sampling techniques, and calibrated classifiers, are employed to rectify this issue [20–25].

Previous research has introduced two types of predictions for asymmetric data: predicting class labels and predicting probabilities. Samplers are used to predict class labels, while calibrators are employed to predict probabilities. Class labels provide a concise description of a data point, accurately representing the actual result of the target variable. Precise labeling of data enhances quality control in ML algorithms, facilitating the training of the model to achieve the desired results. As an alternative to providing a clear class designation, it may be possible to forecast class membership likelihood. This allows users to assess the outcomes within the context of the problem, enabling a forecasting model to distribute the ambiguity of its prediction across various possibilities [26].

Previous studies have highlighted the occurrence of exaggerated and unsatisfactory outputs generated by various ML models. Furthermore, when working with asymmetric data, it has been observed that model outputs tend to vary based on the specific dataset and application. In addition to the existing ML models, new models have been proposed. However, it is noteworthy that none of these studies employed appropriate metrics to evaluate their results. Accuracy was the primary criterion used in all publications, which may not be suitable for dealing with data asymmetries. Interestingly, no articles utilizing hydrological data have been published in this particular research area. Consequently, determining the techniques and metrics that can effectively transform the data based on each model's performance remains a challenge in the context of asymmetric data classification.

This research seeks to rectify these research gaps by leveraging the water potability dataset from the Kaggle repository. Two calibration techniques for probabilistic prediction, five sampling models for label prediction, and six discrete classifiers have been selected to detect and rectify overestimated predictions. The experimental setup consists of three phases. In the initial phase, six traditional ML methods are applied for classification without any sampling to observe the extent of misclassification results when using the water potability dataset.

The experiment is extended in the second phase to assess whether sampling strategies can alleviate the overfitting issues associated with standalone ML classifiers for this specific dataset. In the third phase, this experiment is further extended to examine whether ML classifiers can minimize the loss when combined with calibration techniques across various iterations. Ultimately, various algorithms are analyzed, and the most suitable one is chosen based on the dataset, sampling methods, and effective performance metrics.

The remaining sections of this paper are structured as follows: Section 2 provides a summary of recent research on classifying asymmetric data. Section 3 includes a comprehensive dataset description as well as details on pre-processing procedures. In section 4, traditional ML models, sampling models, and calibration techniques were discussed. Section 5 presents the outcomes and offers metrics-based comparisons with other approaches. The paper concludes in section 6 with discussions on the conclusions and future directions.

2.Literature review

Research papers that focused on the keywords "unbalanced data classification" [27], “class imbalance” [27, 28], "imbalanced classification" [29], “highly imbalanced” [30, 31] were reviewed based on sources from the web. Additionally, research papers that dealt with technical keywords such as "Calibration Techniques," " ML [32]," "DL [33]," and "Sampling Techniques [34, 35]" were also examined. This approach enabled researchers to gain insights into recent strategies established to address this issue. The aforementioned documents served as references (Materials).

Wang et al. [15] presented a personality prediction model that combined particle swarm optimization (PSO) features, synthetic minority oversampling technique (SMOTE), and T-link (Tomek). According to the proposed technique, the mean accuracy for both the unprocessed and processed textual datasets was 75.34% and 78.78%, respectively. Moreover, the mean accuracy for the extended textual dataset was 64.25%, while the shortened textual dataset achieved 75.34% accuracy. It's important to note that this method is most effective for small datasets, which represents the primary limitation of this research.

Joloudari et al. [17] employed various resampling techniques, including random under sampling (RUS), Tomek, one-sided selection, near miss, random over sampling (ROS), and SMOTE. They then utilized DL models for binary data classification. In comparison to the applied sampling strategies, the DL model combined with SMOTE outperformed them, achieving 99.08% accuracy, 99.09% precision, 99.08% sensitivity, 99.09% F1-score, 99.08% G-mean, 99.03% specificity, 99.08% area under the curve(AUC) , and 98.92% kappa. However, one drawback of this research is that it requires a more significant amount of time and computational resources compared to conventional ML techniques.

Zheng et al. [18] introduced an innovative semi-supervised learning-based data pre-processing technique called "near pseudo" (NP). To validate their proposed strategy, experiments were conducted using a state-of-the-art hyper spectral image dataset. The results regarding accuracy rates indicate that NP outperforms other commonly used pre-processing algorithms. When integrated with NP, the classification accuracy of random forest (RF), k-nearest neighbour (KNN), and logistic regression (LR) increased by 1.8%, 4.0%, 6.4%, and 3.7%, respectively. The authors emphasized the importance of carefully selecting the right techniques for transforming data into different feature spaces, stating, "Several methods are capable of converting data into different feature spaces, so selecting the right techniques should be carefully considered.

Werner et al. [21] conducted an evaluation of 9927 papers related to sampling approaches for ML in asymmetrical data scenarios. They conducted this systematic mapping across seven digital libraries. The findings suggest that solutions involving artificial neural networks (ANN) and ensemble ML models tend to have the best performance. According to the authors, using hybrid-sampling strategies in conjunction with ANN and ensemble ML models can yield even better results. Interestingly, none of the 35 studies reviewed utilized synthetic oversampling, which points toward the potential for new pre-processing techniques in this domain.

Swana et al. [23] employed naive bayes (NB), support vector machine (SVM), and KNN to assess classification ability when dealing with asymmetric data. Furthermore, they applied three oversampling techniques, namely SMOTE, Tomek, and SMOTETomek, in conjunction with the aforementioned classifiers to normalize the data and reduce misclassification. Among these samplers, SMOTETomek proved to be the most effective. For both simulated and actual data, NB and KNN outperformed SVM in classification. In particular, KNN combined with SMOTETomek demonstrated superior performance compared to other models.

Bennin et al. [24] evaluated the efficacy of eight data augmentation models for cross-project defect prediction. They employed 34 datasets to test the prognostic capabilities of their model. The authors noted that the selection of source and target data could potentially affect the validity of their research findings. They also acknowledged that it remains uncertain whether their results can be generalized to other datasets that were not utilized as source data. Additionally, the authors reported that the choice of the number of neighboring nations selected for the nearest neighbour (NN) filter influenced the study results.

Devagdorj et al. [28] conducted a comparison of ML models to address class imbalance using smoking cessation data among the Korean population. Initially, they performed feature selection with the assistance of the lasso method and a multicollinearity method. They then employed SMOTE and adaptive

synthetic sampling (ADASYN) resampling techniques to balance the data. Subsequently, gradient boosting tree (GBT), RF, and multi-layer perceptron (MLP) models were utilized for forecasting. The authors emphasized that their forecasting results can vary depending on the data and application.

Johnson and Khoshgoftaar [30] utilized basic methods, namely ROS, RUS, and a hybrid ROS-RUS model, to balance asymmetric network safety data. In their comparison, ROS outperformed RUS and the baseline methods, and ROS-RUS achieved better results than the other two methods. To reduce the risk of misrepresenting the majority class, the authors recommended exploring and integrating techniques for forecasting optimal sample proportions into RUS procedures.

Liang et al. [36] developed the logistic regression-SMOTE (LR-SMOTE) model to resolve data imbalance without generating anomalous data for classification. The authors reported that the newly designed LR-SMOTE outperforms the SMOTE model. It's important to note that the authors' datasets, combined with conventional datasets, had a relatively modest sample size. Therefore, they suggest applying the model to standard high-dimensional data. Hussein et al. [37] proposed a hybrid strategy using an enhanced simulated annealing (SA)-based SVM algorithm and a data preprocessing technique, achieving 89.65% accuracy. Their research focused solely on binary classification, without addressing multi-class classification issues.

Zhao et al. [38] employed cervical cell generation (CCG) - taming transformers to extract a quality dataset from the original unbalanced data. They used the SMOTE-Tomek oversampler to equalize the dataset. The authors acknowledged that this approach introduces technical challenges, making the classification model more complex and potentially increasing the likelihood of classification errors. Although the model demonstrated high accuracy, data imbalance issues persisted. Consequently, they recommended expanding the dataset models and refining the categorization model. Christianto and Rusli [39] developed a student feedback system using recurrent neural networks (RNN) with simple-RNN, long short term memory (LSTM), and gated recurrent unit (GRU) topologies. They applied ROS and SMOTE to address data imbalance. The authors reported that ROS outperformed SMOTE in classification performance. However, they also noted that ROS had a negative impact on LSTM classification performance.

From the review, it is evident that only a few studies have been published on the subject of class asymmetry in hydrological fields. The imbalance between classes poses challenges in obtaining precise probabilities for calibration procedures. Additionally, determining the most effective method for sampling operations is a complex task that has received limited attention from researchers. Moreover, selecting an appropriate metric for the classification of asymmetric data can be quite challenging. Previous studies primarily relied on classification accuracy, which is not well-suited for handling asymmetrical datasets. As a result, our research aims to utilize more suitable measures to identify the superior model for classifying asymmetric data. This research contributes new insights into calibration techniques and sampling methods for datasets with class imbalances, serving as a valuable reference for future research in the field of class imbalance techniques.

3.Materials and methods

3.1Experimental dataset

A publicly available asymmetric hydrological dataset has been 6+- gathered and applied. The water potability dataset is available in a .csv file from Kaggle repository [40]. It comprises 3,277 observations and includes 10 parameters: potential hydrogen (pH), hardness, solids or total dissolved solids (TDS), chloramines, sulphate, conductivity, organic carbon, trihalomethanes (THMs), turbidity, and potability [40–42]. This section delves into the dataset's characteristics. In accordance with the recommendations of the World Health Organisation (WHO) [43], the parameters and their desirable limits are outlined in *Table 1*:

Table 1 Dataset parameters and its desirable limits

S. No.	Parameter	Desirable Limit
1	pH	6.5 – 8.5
2	Hardness	200 mg/L
3	Solids	500 mg/L
4	Chloramines	up to 4 mg/L

S. No.	Parameter	Desirable Limit
5	Sulphate	250-500 mg/L
6	Conductivity	400 μS/cm
7	Organic Carbon	< 2 mg/L
8	THMs	0.06 – 0.2 mg/L
9	Turbidity	1 NTU
10	Potability	0 – Not Potable, 1-Potable

pH is a crucial factor for assessing the acid ratio of water. It indicates whether the water is alkaline or acidic. The primary sources of water hardness are salts composed of calcium and magnesium, originating from geological formations. Water can dissolve various inorganic and organic minerals or salts, including potassium, calcium, sodium, and bicarbonates. These minerals can give the water an undesirable taste and affect its color. Water with a high TDS rating typically has a high mineral content. The two primary chemicals used for water purification purposes in potable water are chlorine and chloramine. Chloramines are commonly produced when ammonia is mixed with chlorine for water purification.

Sulfates are organic compounds naturally present in rocks, soil, and minerals. They find extensive use in the chemical industry for commercial purposes. The electrical conductivity (EC) of water is usually determined by the concentration of dissolved particles. An increase in ion concentration enhances the EC of water. Total organic carbon (TOC) measures the overall quantity of carbon in organic molecules in pure water. Natural organic matter (NOM) from both natural and artificial sources decomposes to form TOC in source waters. Chlorine-treated water may contain chemicals known as THMs. The concentration of THMs in drinking water is influenced by the amount of organic matter in the water, the quantity of chlorine required for purification, and the water's ambient temperature. The turbidity of water is determined by the quantity of suspended solid matter, and it reflects water's ability to transmit light. Water potability, typically expressed on a scale from 0 to 1, indicates its suitability for human consumption [40].

3.2Methodology

In this study, six traditional ML models were employed for both label prediction and probability prediction. These ML models were subjected to three individual and two hybrid re-sampling techniques for label prediction. Subsequently, probability prediction was carried out by combining all classifiers and two calibrators. The performance of the employed models was assessed using a range of evaluation measures, including balanced random accuracy (BRA), sensitivity, specificity, AUC, geometric mean (G-mean), brier score (BS), expected calibration error (ECE), and maximum calibration error (MCE). The results obtained from these evaluation measures, following the sampling and calibration processes, were deemed sufficient to select the top-performing model.

In both prediction stages, nine out of the ten available parameters were utilized as features. The "potability" column (target variable) was employed to predict whether the water is potable or not based on the values of these nine characteristics in the dataset. The feature selection process was conducted using feature statistics as the basis.

The data processing workflow is shown in *Figure 1*. This diagram portrays two types of prediction stages, which are label prediction and probability prediction. The obtained dataset was not in a suitable format and presented difficulties in its usability for constructing ML models. Additionally, it contained missing values and anomalies, rendering it unsuitable for classifier processes. Pre-processing and cleaning of the data are essential steps before utilizing it in ML algorithms. These procedures, such as data cleansing, outlier removal, resampling the data, and formatting, are necessary to transform such data into an appropriate format. Missing values were identified using feature statistics during this phase [44, 45].

Figure 2 depicts the 15% missing data for pH, 24% missing data for sulfate, and 5% missing data for THMs. This issue is resolved by the basic pre-processing techniques through the utilization of simple imputer and feature scaling methods. Secondly, the identification and removal of anomalies within the dataset is a crucial step. To achieve this, the isolation forest (IF) [46] method was utilized to determine the anomaly scores of the experimental data. It is an unsupervised learning technique that builds on the foundation of the DT algorithm [47]. It effectively isolates outliers in the data to identify anomalies. This algorithm scrutinized

the anomaly index of each column in the dataset. *Figure 3* displays the anomaly scores detected for the nine parameters in the dataset. This diagram clearly implies the outlier region and anomaly score for all the parameters employed in the dataset [48]. Further details regarding the six ML techniques, two calibration techniques, and five sampling techniques employed in this study will be provided in the following sections.

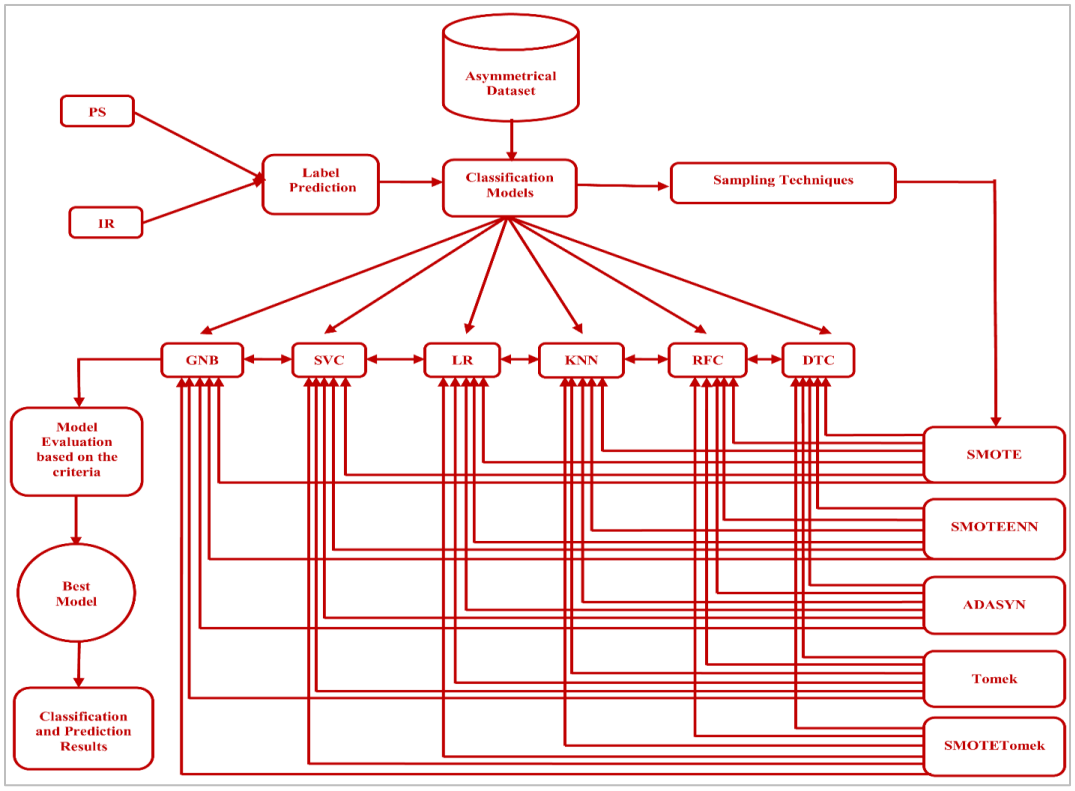


Figure 1 Data processing workflow

	Name	Distribution	Mean	Mode	Median	Dispersion	Min.	Max.	Missing
N	ph		7.05201		7.00604	0.227518	0.00	13.5412	344 (15 %)
N	Hardness		196.207	47.432	196.61	0.168233	47.432	323.124	0 (0 %)
N	Solids		21867.2	320.943	20856.6	0.397774	320.943	61227.2	0 (0 %)
N	Chloramines		7.12698	0.352	7.12674	0.223711	0.352	13.127	0 (0 %)
N	Sulfate		333.144		333.254	0.125242	129	481.031	542 (24 %)
N	Conductivity		425.995	201.62	422.406	0.187805	201.62	708.226	0 (0 %)
N	Organic_carbon		14.3332	2.2	14.2356	0.228152	2.2	28.3	0 (0 %)
N	Trihalomethanes		66.1591		66.4814	0.244131	0.738	124	104 (5 %)
N	Turbidity		3.97134	1.45	3.95365	0.196889	1.45	6.49475	0 (0 %)
C	Potability			0		0.668			0 (0 %)

Figure 2 Feature statistics

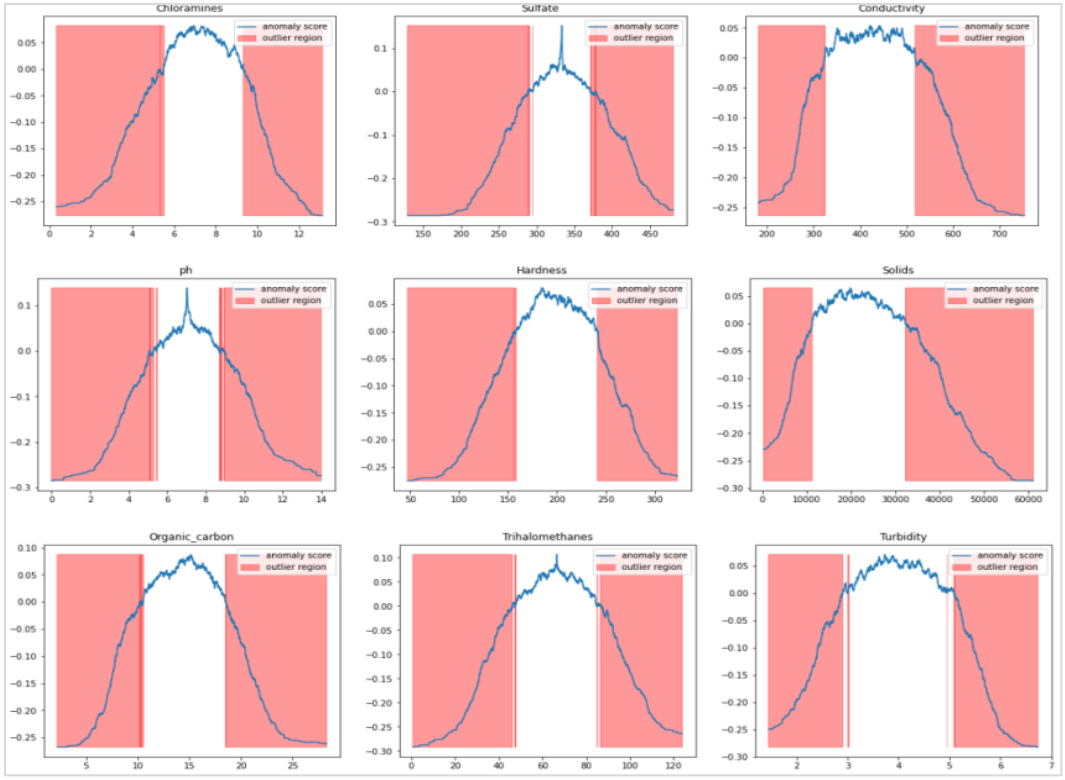


Figure 3 Anomaly score calculation by IF mechanism

3.3ML models

Numerous methods were developed using ML to deal the challenges posed by asymmetric data. These approaches typically categorized into two groups: algorithmic or internal-level approaches and data-level approaches for handling asymmetric data. In this research, we focused on resolving the issue of data asymmetry through an algorithm-level approach. We employed various ML techniques in the current study, including support vector classifier (SVC), KNN, gaussian naive bayes (GNB), RF, LR, and DT [8, 13].

3.3.1SVC

SVC effectively solved both linear and nonlinear issues [20]. It operated as a rapid 2-class group classification algorithm in supervised learning [44]. When provided with two coordinates as input, it produced an outcome in the form of a hyperplane model. The concept behind utilizing the SVC algorithm for classification involved locating the optimum hyperplane and the absolute optimum division space in distinct classes within the input space. However, our research findings revealed that it did not provide suitable results for cases like probability prediction and binary classification. This algorithm excelled primarily in multi-class classification.

3.3.2KNN

KNN served as a nonlinear regression approach in ML and was also utilized for classification. It generated columns with features obtained from the target values of the dataset and used these values to determine similarities between the validation data and the prediction data. The target value's role was to forecast performance based on these similarities [37]. Since all processing occurred concurrently during testing and involved an iterative process of training samples, it calculated comparable values each time to form the clustering results [47]. However, it was impractical for large datasets and was characterized as a time-consuming, or "slothful," learner algorithm.

3.3.3GNB

GNB was a likelihood-based algorithm utilized to classify binary and multiclass data. It was a quickand traditional ML technique that relied on the Bayes theorem and probability theory as its foundation for predicting likelihood. One advantage of this approach was its swift evaluation for time-consuming calculations [46, 48]. The term "naïve" referred to the assumption that the model's components were independent. GNB's concept of using a logarithmic transformation of probabilities [49] to prevent underflow was another advantage of employing this method. It provided an excellent solution for

classification issues that required likelihood estimations [50].

3.3.4RF

RF was predominantly used to address classification and regression problems. As the name suggests, a forest comprises trees, and a healthier forest contains more trees. Similarly, the RF method constructed decision tree (DTs) from data samples, extracted predictions from each tree, and then voted on the best option. This ensemble method outperformed a single DT by averaging the results to reduce over fitting [51, 52]. Owing to the randomized and decorrelated nature of RF, it had the capacity to establish associations or links between input and output parameters even when their relationship was complex and nonlinear [53].

3.3.5LR

LR, recognized for its sigmoid function, was used to compute or predict the probability of a binary outcome. The result of this technique was binary, making decisions based on 'true' or 'false,' 'yes' or 'no' [52]. It exhibited remarkable effectiveness depending on the data or application [53]. Rapid learning instances combined with low computational resource requirements allowed it to easily scale, even with large datasets.

3.3.6DT

The DT was the most effective method in ML and a knowledge representation technique for data. It offered technologies for analysing vast, intricate amounts of data to discover insightful patterns [54]. Data scientists found DTs to be a valuable data extraction technique for anticipating asymmetric data based on water quality factors, aiding in determining which data should progress to the next prediction stage.

3.4Sampling models

To mitigate the adverse effects of an asymmetrical dataset, sampling approaches are commonly employed within the realm of ML models. The sampling process is typically distinguished into three main categories [23], encompassing under-sampling, over-sampling, and threshold moving [24]. In this study, two over-sampling methods, one under-sampling method, and two hybrid sampling techniques were utilized. The various data sampling techniques employed in the current investigation are as follows: SMOTE, ADASYN, Tomek, SMOTETomek, and SMOTE with edited nearest neighbour (SMOTEENN) [23–29].

3.4.1SMOTE

This method was devised to mitigate the impact of class asymmetry on prediction accuracy [23]. It has gained recognition and is widely employed in published literature for various predictive research endeavours. SMOTE leverages each minority class data point to create "synthetic" minority samples [24]. It employs KNN algorithm to locate neighbouring points for each minority class sample, subsequently selecting the kth neighbor at random to generate a new synthetic sample.

3.4.2ADASYN

The ADASYN learning algorithm primarily focuses on the challenging-to-learn instances within the minority class data [20]. It produces a variable for the sample count based on an examination of the internal dispensation of the oversampled class. Its key advantage lies in preventing the duplication of minority data [24]. Unlike the SMOTE method, it produces non-identical synthetic data for each minority class instance, emphasizing the challenging cases within the minority class instances [28].

3.4.3Tomek

Tomek is an under-sampling technique aimed at the elimination of redundancies. One such under-sampling technique is an adaptation of the condensed nearest neighbour (CNN) rule, referred to as the Tomek links approach [17]. This technique is used to identify whether a Tomek link can be established between different sets of samples. Such links have the significant feature of exclusively removing undesirable samples [23, 35]. A selection of class labels [55, 56] identified as Tomek links can be eliminated, which is highly useful for identifying samples from different classes.

3.4.4SMOTEENN

SMOTEENN was created by combining the over-sampling method SMOTE [14] and the under-sampling method edited nearest neighbour (ENN) rule [29]. This algorithm significantly improved sensitivity and specificity [36]. It serves as an effective solution to address the shortcomings of both the SMOTE and ENN techniques [57–59].

3.4.5SMOTETomek

SMOTETomek, a hybrid approach, was developed by merging the over-sampling SMOTE and under-sampling Tomek methods [34], hence the name SMOTETomek [35, 38]. The algorithmic workflow of the SMOTETomek method combines SMOTE with the Tomek link approach, effectively creating a pipeline [23]. This approach offers a compelling solution to mitigate the drawbacks associated with both the SMOTE and Tomek link techniques [24].

3.5Calibration models

Calibration is a method employed to acquire precise probability estimates for practical applications of classification issues [16]. In practical settings, the

number of training samples per class often varies, making it crucial to address data asymmetry. The use of calibration techniques is an effective strategy for mitigating the impact of asymmetric data. This study utilized two calibration methods, isotonic regression (IR) and platt scaling (PS), to predict probabilities [25]. The two calibration procedures mentioned above consist of one non-parametric approach and one parametric approach.

3.5.1IR calibration

IR calibration is a non-parametric regression method [49]. Non-parametric implies that no inferences are made about variables such as constant interpolation, variance, or shape. When addressing a calibration issue, this method seeks to perform regression on the initial calibration curve. It allows for arbitrary shaping without presuming the form of the target value to address the asymmetric nature of the ML model. This method particularly excels with large datasets.

3.5.2PS calibration

PS calibration is a parametric approach [49]. Initially designed for calibrating SVM model, it is now applied to other classifications as well. SVMs can only produce results on the samples based on the predicted edges because they are optimized using hinge loss. To address this limitation, John Platt proposed the use of PS in combination with LR to convert results into probability estimates [50, 60]. The sigmoid function assigns probability values to discrete classes (0 and 1) [50, 61]. Its probabilistic nature makes it suitable for the current water quality prediction.

4.Results

The outcomes of the proposed methodology are presented in this section, which is divided into two parts: label prediction and probability prediction. In the first phase, standalone ML classifiers were used to identify classification results that might be overvalued. The second phase aimed to identify an appropriate sampling model using multiple sampling procedures. Finally, through various iterations, an adaptive calibration mechanism for probabilistic prediction was discovered. The entire experimental investigation was conducted using the Anaconda3 2020.11 (Python 3.8.5 64-bit) platform and the imbalanced learning (imblearn) Package.

4.1Performance evaluation

Evaluation metrics played a crucial role in assessing classification efficacy and guiding classifier modeling. When dealing with asymmetric data classification, different evaluation metrics were essential. While accuracy is a common metric for classification [60–62], it is not suitable for asymmetric classification [63] because a less effective model can achieve a higher accuracy. Evaluation of expected and predicted class labels or assessment of probabilities for the anticipated class labels were needed for classification issues.

4.1.1Metrics for label prediction

In the context of asymmetric data classification, BRA and the G-mean metrics are considered the most reliable performance indicators for classification algorithms in label prediction. Additionally, the AUC analysis was employed to visualize the categorization of the dataset used for label prediction. Equations 1 to 5 provide the mathematical expressions for calculating specificity, sensitivity, BRA, G-mean, and AUC, as follows:

Specificity: The percentage of accurately detected negatives over all possible negative forecasts produced by the algorithm is measured by specificity [63]. It is also referred to as the true negative rate.

$$Specificity = \frac{T_N}{F_P + T_N} \quad (1)$$

The true positive, true negative, false positive, and false negative are denoted as TP, TN, FP, and FN respectively.

Sensitivity: A performance metric derived from the positive observations is sensitivity or recall [63]. The percentage of positive observations that were identified accurately as positive is displayed by sensitivity.

$$Sensitivity = \frac{T_P}{T_P + T_N} \quad (2)$$

The true positive, true negative is denoted as TP and TN respectively.

BRA: The efficiency of BRA's uses outweighs the normal classification accuracy. So, Dealing with asymmetric data required balanced accuracy for the most of the binary and multi-class categorization. It is the mathematical average of sensitivity and specificity [64].

$$BRA = \frac{Specificity + Sensitivity}{2} \quad (3)$$

G-mean: The evaluation measure G-mean is frequently employed in asymmetric data analysis [65].The G-mean of sensitivity and specificity is denoted by G-mean [64].

$$G - mean = sqrt\ (Sensitivity \times Specificity) \quad (4)$$

AU-ROC curve: To assess how successfully a classifier balances out its TP rates and FP rates, AUC

provides a value representing a scalar [66].This measure's closest representation is given below

$$AUC = \frac{1+T_P-F_P}{2} \quad (5)$$

The true positive, false positive is denoted as TP and FP respectively.

4.1.2Metrics for probability prediction

For probability prediction in classification algorithms, BS, ECE, and MCE are considered the most reliable performance indicators. Additionally, the calibration curve, also known as the reliability curve, was utilized to visualize the categorization of the dataset used for probability prediction [17, 37]. Equations 6 to 8 provide the mathematical expressions for calculating BS or log-loss, ECE, and MCE, as follows:

BS: BS, referred to as log-loss values and mean squared error (MSE). It is a well-liked statistical-based evaluation metric for assessing how well the likelihood estimator performs [61].It also measures how close the calibrated probabilities are to 0 or 1and describes how closely the calibrated probabilities resemble the real probabilities.

$$BS = \frac{1}{N}\sum_{i=1}^{b} N_i(f_i - o_i)^2 + \frac{1}{N}\sum_{i=1}^{b}(o_i(1 - o_i)) \quad (6)$$

Where N_i is the total number of occurrences in the i^{th} bin, f_i denotes the percentage of positive occurrences, and e_i denotes the mean calibrated probability in the i^{th} bin [16].

ECE: The ECE assesses the effectiveness of the calibration overall [16].The calibrated likelihoods must be ordered and split into various bins before being used to compute ECE.

$$ECE = \sum_{i=1}^{b} \pi_i . |o_i - f_i| \quad (7)$$

If fi is the mean calibrated probability in that bin, o_i is the proportion of positive instances in the i^{th}bin, and i is the proportion of occurrences that fit in the i^{th} bin.

MCE: The sustainability of calibration is assessed by utilizing the MCE. A calibration method's MCE value will be lower compared to other methods if it is more reliable and consistent [16]. Consequently, it is essential to evaluate the sustainability of calibration.

$$MCE = max_{i=1}^{b} |o_i - f_i| \quad (8)$$

Calibration curve: Calibration curves (reliability diagrams) were used to predict the likelihood of a given group utilizing classifiers and to determine outcomes [66, 67]. It provides a predictive means of determining if the scores are reliable.

4.2 Label prediction results

This section presents the experiments and simulation findings for the proposed strategy aimed at addressing unbalanced hydrological data classification and prediction estimation. The comparative outcomes are displayed in this section, utilizing evaluation metrics such as BRA, sensitivity, specificity, AUC, and G-Mean. The results are given in *Table 2* for six classification models: LR, SVM, GNB, KNN, DT, and RF in combination with five samplers: SMOTE, ADASYN, Tomek, SMOTEENN, and SMOTETomek, respectively.

Table 2 Comparative analyses of various classification methods in combination with samplers for Label Prediction

Models	Sensitivity (%)	Specificity (%)	BRA (%)	G-mean (%)	AUC (%)
LR	80.01	84.21	82.06	81.98	0.824
LR+SMOTE	75.10	79.09	77.31	76.97	0.777
LR+TOMEK	68.05	72.20	70.03	69.97	0.698
LR+ADASYN	72.12	80.62	76.28	75.89	0.765
LR+SMOTEENN	74.39	78.96	76.91	75.97	0.758
LR+SMOTETOMEK	68.08	84.03	76.24	75.58	0.765
SVM	96.01	100	98.19	97.98	0.983
SVM+SMOTE	91.50	95.90	93.03	92.98	0.934
SVM+TOMEK	94.23	98.17	96.67	95.98	0.961
SVM+ADASYN	98.11	98.00	97.00	96.99	0.960
SVM+SMOTEENN	**95.50**	**98.50**	**97.01**	**96.99**	**0.968**
SVM+SMOTETOMEK	95.51	96.52	96.63	95.10	0.960
GNB	90.42	86.13	88.90	87.98	0.888
GNB+SMOTE	80.09	90.34	85.11	84.85	0.855
GNB+TOMEK	82.12	86.00	84.66	83.98	0.842
GNB+ADASYN	87.90	91.93	89.23	88.98	0.892
GNB+SMOTEENN	85.51	92.50	89.89	88.93	0.892
GNB+SMOTETOMEK	84.02	94.14	89.82	88.86	0.887
KNN	85.13	89.90	87.70	86.98	0.874

Models	Sensitivity (%)	Specificity (%)	BRA (%)	G-mean (%)	AUC (%)
KNN+SMOTE	82.92	84.86	83.94	82.99	0.831
KNN+TOMEK	68.96	76.23	72.88	71.89	0.725
KNN+ADASYN	76.56	80.34	78.80	77.97	0.770
KNN+SMOTEENN	79.09	83.45	81.67	80.98	0.812
KNN+SMOTETOMEK	80.31	86.90	83.56	82.95	0.826
DT	98.76	100	99.80	98.99	0.992
DT+SMOTE	92.34	96.21	94.92	93.98	0.941
DT+TOMEK	90.43	94.22	92.45	91.98	0.916
DT+ADASYN	88.65	92.21	90.87	89.98	0.899
DT+SMOTEENN	94.23	96.34	95.78	94.99	0.948
DT+SMOTETOMEK	94.50	97.51	96.34	95.99	0.956
RF	98.22	99.11	98.54	98.50	0.984
RF+SMOTE	94.09	98.08	96.15	95.98	0.958
RF+TOMEK	95.16	99.22	97.32	96.98	0.965
RF+ADASYN	96.31	98.55	97.44	96.99	0.970
RF+SMOTEENN	**98.02**	**99.01**	**98.07**	**98.50**	**0.976**
RF+SMOTETOMEK	97.32	99.00	97.11	97.99	0.967

According to *Table 2*, it is observed that the RF+SMOTEENN model has the best performance with 98.07% BRA, 98.02% sensitivity, 99.01% specificity, 0.976% AUC, and 98.50% G-Mean. Following that, the SVM+SMOTEENN model demonstrates better performance with 97.01% BRA, 95.50% sensitivity, 98.50% specificity, 0.968% AUC, and 96.99% G-Mean. The best outcomes are highlighted in red and in bold.

Additionally, *Figure 4* in our experiment displays ROC plots based on the Top AUC scores attained using the ML models. *Figure 4(a)* depicts an ROC plot in relation to different classifiers before sampling. This plot indicates that, when comparing the standalone classifiers, all of them exhibit over-fitting results before sampling. In particular, it reveals that DT produces exaggerated results when compared to other discrete classifiers.

Figure 4(b) portrays a ROC plot in relation to different classifiers with SMOTE sampling. When comparing the classifiers, this plot shows that RF has the highest AUC, while LR has the lowest AUC for the dataset under consideration.

Figure 4 (c) depicts a ROC plot in relation to different classifiers with Tomek sampling. When comparing the classifiers, this plot illustrates that RF and SVC have the highest AUC, whereas LR has the lowest AUC for the dataset under consideration.

Figure 4 (d) shows a ROC plot in relation to different classifiers with adasyn sampling. When comparing the classifiers, this plot demonstrates that RF has the highest AUC and LR has the lowest AUC for the dataset under consideration.

Figure 4 (e) depicts a ROC plot in relation to different classifiers with SMOTEENN sampling. When comparing the classifiers, this plot displays that RF has the highest AUC and LR has the lowest AUC for the dataset under consideration.

Figure 4 (f) portrays a ROC plot in relation to different classifiers with SMOTETomek sampling. When comparing the classifiers, this plot indicates that RF has the highest AUC, while LR has the lowest AUC for the dataset under consideration.

From overall analysis, hybrid samplers in combination with all classifiers are better than other samplers' combinations continuously. Significantly, RF and SVM in combination with hybrid samplers providethe top AUC scores. More importantly, *Figure 4(e)* and *Table 2* demonstrate that the RF+SMOTEENN model achieves the highest AUC value.

4.3Probability prediction results

The experimental data divided into three subsets: 20% for testing, 60% for training, and 20% for calibration or validation, following the recommendation in [16]. This division was essential for validating the model's performance. In the initial stage, we scrutinized two chosen probabilistic and six ML algorithms. To assess the discrimination and calibration capabilities of the model, we employed measurements such as the BS, ECE and MCE [46]. Before calculating the ECE, we sorted the calibrated probabilities and divided them into various bins. In

this study, we randomly divided the bins into three categories: 10, 100, and 1000. We utilized these specified average probability ranges (bins) to define the fraction of positives [68].The results illustrated that IR outperformed other techniques on all six classifiers. Notably, IR exhibited superior performance when applied with GNB and LR as the classifiers. As per the data in *Table 3*, it is evident that the GNB + IR model exhibits best performance with lower error rate in each bin compared with other algorithms. Our results differed based on bin frequency when using experimental data. The probability prediction results for these segmented bins are shown in *Table 3*.

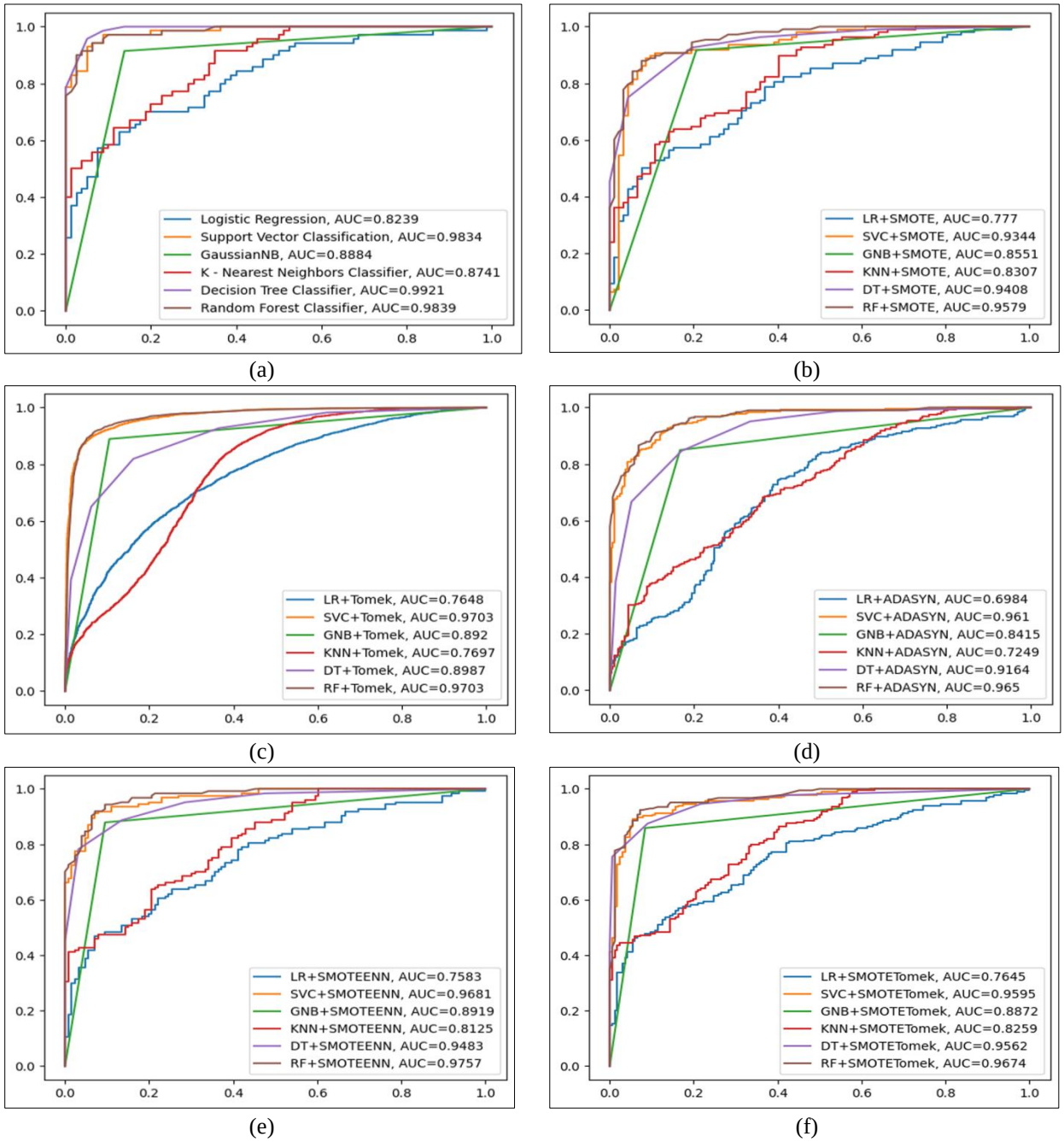


Figure 4 AUROC plots for the models from (a-f)

Table 3 Error metrics for probability prediction by using classifiers with calibrators

Bins	Techniques	BS		ECE		MCE	
		IR	PS	IR	PS	IR	PS
b = 10	GNB	**0.227**	0.228	**0.231**	0.323	**0.336**	0.991
	LR	0.660	0.686	0.668	0.790	0.702	1.524
	KNN	0.804	0.867	0.903	1.429	1.226	2.197
	RF	0.642	0.701	0.698	2.275	2.008	5.714
	DT	0.961	1.957	1.898	2.117	2.008	2.795
	SVC	0.660	0.719	0.872	2.938	2.713	4.005
b = 100	GNB	**0.234**	0.235	**0.240**	0.345	**0.355**	0.999
	LR	0.669	0.690	0.673	0.800	0.711	1.541
	KNN	0.874	0.872	0.931	1.444	1.232	2.199
	RF	0.652	0.713	0.709	2.283	2.019	5.721
	DT	0.983	1.967	1.918	2.131	2.052	2.798
	SVC	0.680	0.731	0.880	2.945	2.725	4.012
b = 1000	GNB	**0.238**	0.238	**0.243**	0.353	**0.358**	1.056
	LR	0.671	0.699	0.677	0.815	0.715	1.579
	KNN	0.877	0.878	0.938	1.458	1.238	2.201
	RF	0.656	0.725	0.712	2.290	2.029	5.730
	DT	0.985	1.979	1.921	2.141	2.064	2.802
	SVC	0.683	0.744	0.885	2.949	2.732	4.039

Overall analysis shows an empirical assessment of the likelihood of positive group occurrences for the group under discussion. In a model that has been properly calibrated, the likelihood of positive group events occurring in a specific bin for the class contemplated corresponds to the average forecasted likelihood.

Figure 5 (a-f) illustrates linear correlation connecting the average anticipated likelihood and the likelihood of the positive group occurrence appearing in that bin for the group under consideration.

Figure 5 (a) portrays the calibration plot (reliability curve) based on the error metrics attained using the GNB model with calibrated classifiers. When using the GNB, IR achieves excellent calibration, as shown in this plot. Simultaneously, the PS and GNB arcs are plotted beneath the diagonal, indicating that the algorithm has over-fitted, and the likelihoods are excessively high.

Figure 5 (b) portrays the reliability curve based on the error metrics attained using the LR model with calibrated classifiers. Based on the simulation measures and the curve, LR+IR are ranked second and exhibits marginally inferior performance than GNB+IR. LR with PS produces arcs that are similar to IR.

Figure 5 (c, f) interprets the reliability curve based on the error metrics attained using the KNN and SVC models with calibrated classifiers. The presence of an IR arc above the diagonal indicates that the algorithm has under fitted, and the likelihoods are excessively tiny.

Figure 5 (d) portrays the reliability curve based on the error metrics attained using the RF with calibrated classifiers. The presence of an IR arc is plotted between the diagonal from bottom to top. Simultaneously, the PS and GNB arcs are plotted beneath the diagonal, indicating that the algorithm has over-fitted and the likelihoods are excessively high.

Figure 5 (e) interprets the reliability curve based on the error metrics attained using the DT with calibrated classifiers. The presence of an IR arc on the diagonal suggests that the algorithm is perfectly calibrated. The presence of the PS arc below the diagonal indicates that the algorithm has over fitted, and the likelihoods are excessively high.

If the ML algorithm predicts correctly, the proportion of prominent group classifications and the average likelihood allocated to the most prevalent classes in every bin should be near to one another. Failure to act precisely will cause these two values to be in different locations. As per the observation, it was concluded that the best performing model for calibration is GNB+IR.

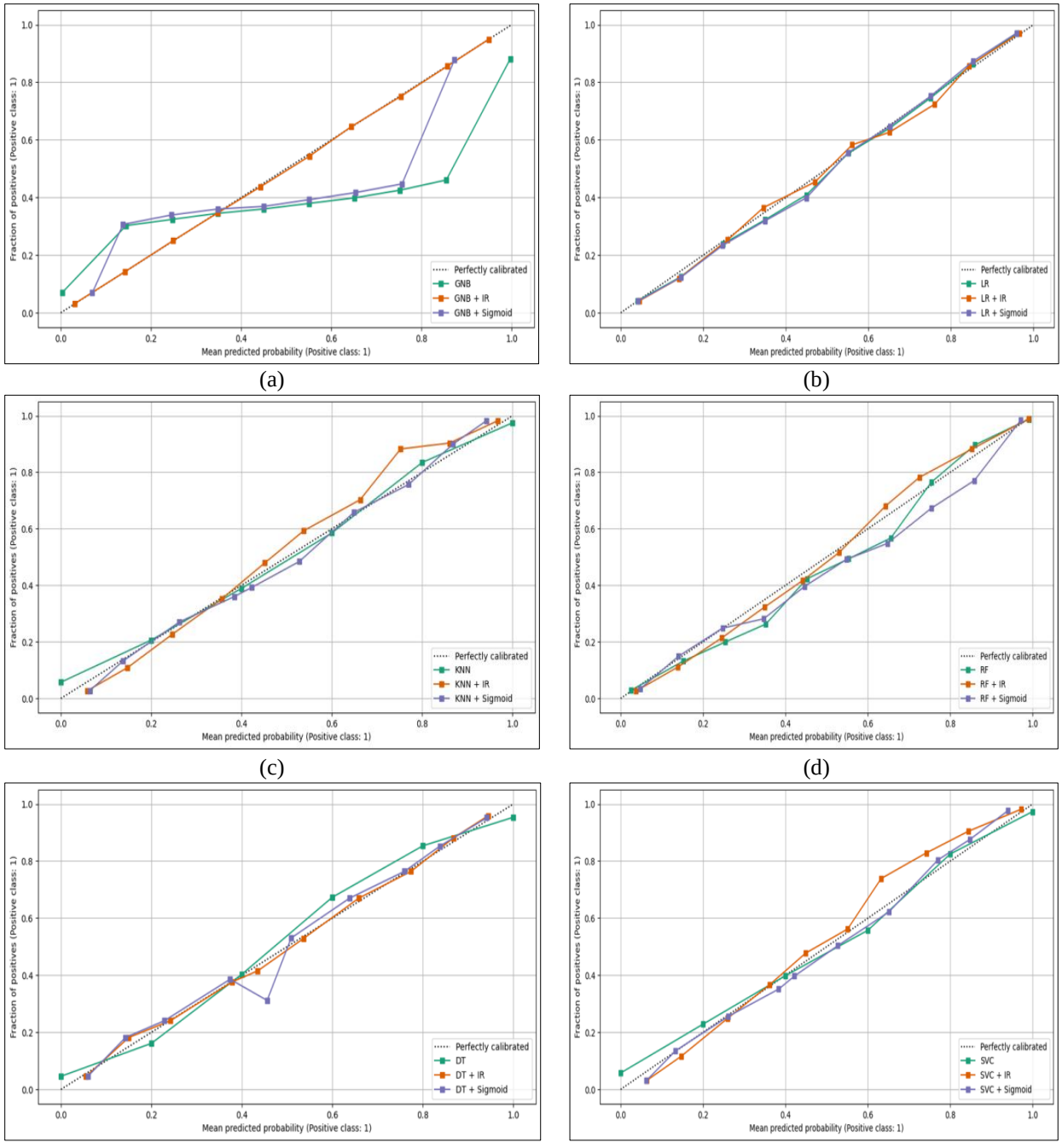


Figure 5 Calibration plots for the models from (a-f)

5.Discussion

This study focuses on two types of predictions using sampling and calibration techniques with various classifiers to address data asymmetry issues. Asymmetric datasets were obtained from the Kaggle repository for classifier performance evaluation. The initial data pre-processing stage involves dataset cleansing to handle missing values, followed by feature selection to facilitate data splitting and target identification. The two prediction types explored in this study are label prediction and probability prediction.

Label prediction aims to identify a suitable sampling model. Initially, we employ well-known classifiers, including RF, GNB, KNN, DT, LR, and SVC, for classification without sampling. We assess performance using measures such as BR, sensitivity, specificity, AUC, and G-Mean. Subsequently, we apply various sampling approaches, such as SMOTE, Tomek Links, ADASYN, SMOTEENN, and SMOTETomek, to transform the data. We then reapply the same ML techniques to compute the metrics.

The results showed that individual classifiers yielded exaggerated results before sampling, while ML classifiers demonstrated improved performance after sampling. Hybrid sampling methods outperformed other approaches with all classifiers. Based on the simulation measures, the combination of RF and SMOTEENN performed exceptionally well, achieving a 98.07% BRA, 98.02% sensitivity, 99.01% specificity, 0.976% AUC, and 98.50% G-Mean. SMOTEENN emerged as an effective model for label prediction.

Following label prediction, probability prediction aims to find an adaptive calibration model. To address data asymmetry challenges in probability prediction, we introduce two widely used calibration approaches, IR and PS. We randomly divide the dataset into three subsets: training, calibration, and testing sets, as recommended in [16], to evaluate calibration performance.

The classification model is trained using the training set, while the calibration model is trained using the calibration set. We build ML models and apply them to calibrate results using the validation set. Subsequently, we assess the efficacy of each calibration method using the testing set. We sort the calibrated probabilities and divide them into bins, with bin sizes randomly selected from three options: 10, 100, and 1000.

Compute various metrics, including BS, ECE, MCE, and reliability curves, to assess the performance of each ML model. IR consistently outperforms the PS technique when combined with all six classifiers, with the best results achieved by GNB in combination with IR, displaying the lowest error rate in each bin. IR emerges as an effective model for probability prediction. However, it's important to note that the error rate increases as the bin size increases. Therefore, based on our analysis, label prediction is found to be more suitable than probability prediction.

Variations in model performance are observed in both types of predictions. Some shortcomings are identified with classifiers when dealing with asymmetric datasets. The limitations of each technique are discussed in this section. For categorical predictions, LR is effective but requires independence of every parameter in the data sample. GNB makes assumptions about sample distribution. KNN has limitations related to data storage for large search problems.

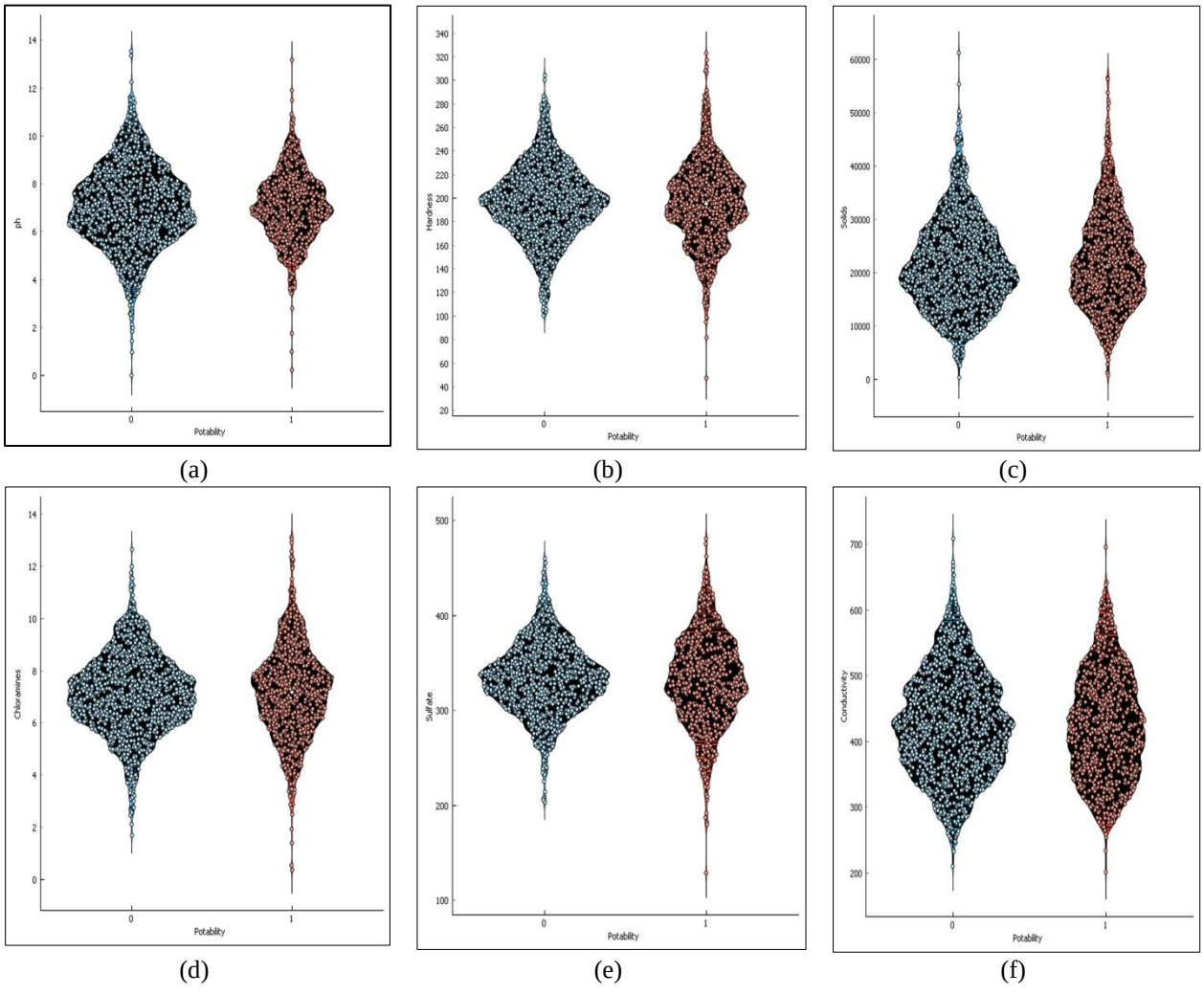
SVM faces challenges due to the lack of transparency caused by high-dimensional data. While DT is quick for learning and prediction, it is sensitive to minor dataset changes and prone to overfitting. In contrast, RF learns quickly and produces effective forecasts once trained, making it a recommended model for forecasting. In this research, the RF method is applied to improve water potability state prediction.
In *Figure 6 (a-i)*, the RF algorithm's prediction results are depicted. Water potability was predicted based on the levels of nine parameters (pH, Hardness, solids, chloramines, sulfate, conductivity, organic_carbon, THMs, and turbidity). As per the dataset description, a label of zero signifies "not drinkable," while a label of one indicates "potable." It's evident that all factors are predicted to be higher in the "not drinkable" category compared to the "potable" category.

A complete list of abbreviations is shown in *Appendix I.*

6.Conclusion and future direction

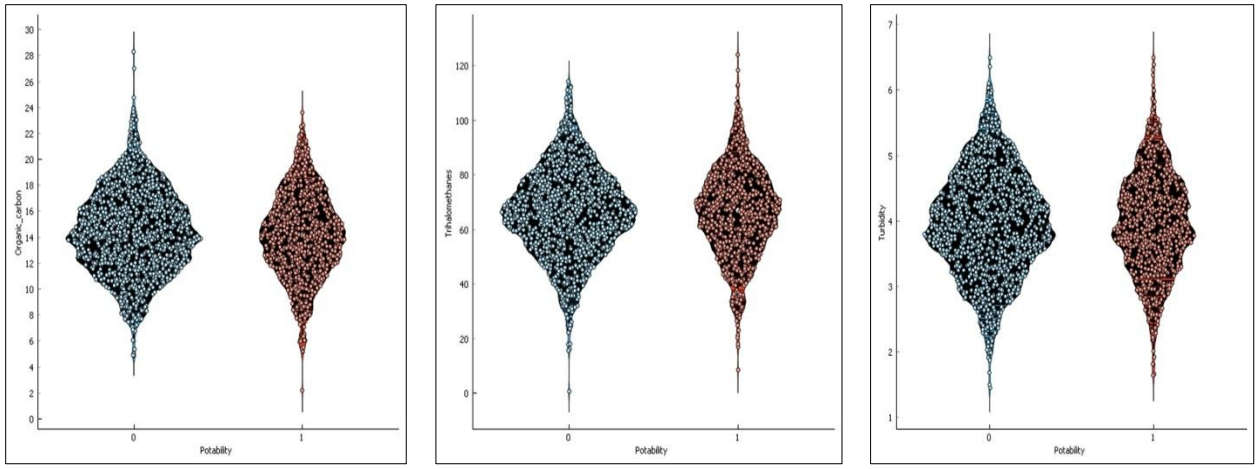
The primary goal of this study was to determine the most effective approaches for label and probability prediction when dealing with asymmetric data. To achieve this, we thoroughly investigated various ML classifiers and their combinations with different resampling techniques for the classification and prediction of asymmetric data. Additionally, we explored ML classifiers, including IR, PS, and the integration of these calibration techniques for probabilistic forecasting. The dataset employed in this research originates from recorded physiological and biochemical properties of water, reflecting real-world applications. By conducting rigorous comparisons, we have assessed the performance of these methods to identify the most suitable strategies for classifying and predicting asymmetric data. Our findings offer clear recommendations for addressing

these challenges effectively. For label prediction tasks, the SMOTEENN emerged as the most reliable choice. This approach consistently demonstrated robust performance, significantly enhancing the accuracy and reliability of classification results. In the domain of probability prediction, the IR calibration method stood out as the superior option. It effectively mitigated issues such as over-fitting, under-fitting, and misclassification by ensuring well-calibrated probability estimates. This research contributes not only to the identification of optimal strategies for asymmetric data classification and prediction but also to addressing common issues such as over-fitting and misclassification. By integrating leading ML models with appropriate samplers and calibrators, effective solutions for these critical challenges in real-world applications are provided. Discovering and establishing a standard for environmental modelling is a never-ending hard slog. The findings of this study are influenced by the dataset used. This cannot be stated to apply to other dataset depending on its features, though. Consequently, it is recommended to broaden the scope of this research by utilizing different datasets. Meanwhile, it is believed that the study will be more effective if this work is extended to use more calibration techniques in research. As researchers seek to select the best model for asymmetric data, the recommendations outlined above are expedient.



(a) (b) (c)

(d) (e) (f)



(g) (h) (i)

Figure 6 Water potability prediction based on the dataset parameters (a-i)

Acknowledgment

This work has been financially supported by the RUSA – Phase 2.0 grant, as sanctioned in Letter No. F. 24-51/2014-U, Policy (TNMulti-Gen), Department of Education, Government of India, dated October 9, 2018.

Conflicts of interest

The authors have no conflicts of interest to declare.

Author's contribution statements

C. Kaleeswari: Data collection, writing – original draft, model training, analysis and interpretation of results. **K. Kuppusamy:** Conceptualization, manuscript preparation - writing – review and editing, supervision. **A. Senthilrajan:** Supervision, design of the work, investigation on the result analysis.

References

[1] Tazoe H. Water quality monitoring. Analytical Sciences. 2023;39(1):1-3.

[2] Adeleke IA, Nwulu NI, Ogbolumani OA. A hybrid machine learning and embedded IoT-based water quality monitoring system. Internet of Things. 2023; 22:100774.

[3] Wang Z, Jia D, Song S, Sun J. Assessments of surface water quality through the use of multivariate statistical techniques: a case study for the watershed of the Yuqiao reservoir, China. Frontiers in Environmental Science. 2023; 11:1-15.

[4] Banerjee P, Dehnbostel FO, Preissner R. Prediction is a balancing act: importance of sampling methods to balance sensitivity and specificity of predictive models based on imbalanced chemical data sets. Frontiers in Chemistry. 2018; 6:1-11.

[5] Chinnakkaruppan K, Krishnamoorthy K, Agniraj S. A hybrid approach for forecasting the technical anomalies in sensor-based water quality distribution data. In international conference on power, instrumentation, energy and control 2023 (pp. 1-5). IEEE.

[6] Piao C, Wang N, Yuan C. Rebalance weights adaboost-SVM model for imbalanced data. Computational Intelligence and Neuroscience. 2023; 2023:1-26.

[7] Pandey S, Kumar K. Software fault prediction for imbalanced data: a survey on recent developments. Procedia Computer Science. 2023; 218:1815-24.

[8] Douzas G, Bacao F, Fonseca J, Khudinyan M. Imbalanced learning in land cover classification: improving minority classes' prediction accuracy using the geometric SMOTE algorithm. Remote Sensing. 2019; 11(24):1-14.

[9] Liang Z, Wang H, Yang K, Shi Y. Adaptive fusion based method for imbalanced data classification. Frontiers in Neurorobotics. 2022; 16:1-8.

[10] Ahmed J, Green II RC. Predicting severely imbalanced data disk drive failures with machine learning models. Machine Learning with Applications. 2022; 9:1-12.

[11] Basora L, Bry P, Olive X, Freeman F. Aircraft fleet health monitoring with anomaly detection techniques. Aerospace. 2021; 8(4):1-33.

[12] Muharemi F, Logofătu D, Leon F. Machine learning approaches for anomaly detection of water quality on a real-world data set. Journal of Information and Telecommunication. 2019; 3(3):294-307.

[13] Bao F, Wu Y, Li Z, Li Y, Liu L, Chen G. Effect improved for high-dimensional and unbalanced data anomaly detection model based on KNN-SMOTE-LSTM. Complexity. 2020; 2020:1-7.

[14] Muntasir NM, Faisal F, Jahan RI, Al-monsur A, Ar-rafi AM, Nasrullah SM, et al. A comprehensive investigation of the performances of different machine learning classifiers with SMOTE-ENN oversampling technique and hyperparameter optimization for imbalanced heart failure dataset. Scientific Programming. 2022; 2022:1-7.

[15] Wang ZH, Wu C, Zheng K, Niu X, Wang X. SMOTETomek-based resampling for personality recognition. IEEE Access. 2019; 7:129678-89.
[16] Huang L, Zhao J, Zhu B, Chen H, Broucke SV. An experimental investigation of calibration techniques for imbalanced data. IEEE Access. 2020; 8:127343-52.
[17] Joloudari JH, Marefat A, Nematollahi MA, Oyelere SS, Hussain S. Effective class-imbalance learning based on SMOTE and convolutional neural networks. Applied Sciences. 2023; 13(6):1-34.
[18] Zheng X, Jia J, Chen J, Guo S, Sun L, Zhou C, et al. Hyperspectral image classification with imbalanced data based on semi-supervised learning. Applied Sciences. 2022; 12(8):1-19.
[19] Schmidt L, Heße F, Attinger S, Kumar R. Challenges in applying machine learning models for hydrological inference: a case study for flooding events across Germany. Water Resources Research. 2020; 56(5):1-10.
[20] Rahman AA, Prasetiyowati SS, Sibaroni Y. Performance analysis of the imbalanced data method on increasing the classification accuracy of the machine learning hybrid method. JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika). 2023; 8(1):115-26.
[21] Werner DVV, Schneider AJA, Dos SCR, Da SPPR, Victória BJL. Imbalanced data preprocessing techniques for machine learning: a systematic mapping study. Knowledge and Information Systems. 2023; 65(1):31-57.
[22] Thabtah F, Hammoud S, Kamalov F, Gonsalves A. Data imbalance in classification: experimental evaluation. Information Sciences. 2020; 513:429-41.
[23] Swana EF, Doorsamy W, Bokoro P. Tomek link and SMOTE approaches for machine fault classification with an imbalanced dataset. Sensors. 2022; 22(9):1-21.
[24] Bennin KE, Tahir A, Macdonell SG, Börstler J. An empirical study on the effectiveness of data resampling approaches for cross-project software defect prediction. IET Software. 2022; 16(2):185-99.
[25] Aggarwal U, Popescu A, Belouadah E, Hudelot C. A comparative study of calibration methods for imbalanced class incremental learning. Multimedia Tools and Applications. 2022:1-20.
[26] Kuhn M, Johnson K, Kuhn M, Johnson K. Remedies for severe class imbalance. Applied Predictive Modeling. 2013:419-43.
[27] Liu L, Wu X, Li S, Li Y, Tan S, Bai Y. Solving the class imbalance problem using ensemble algorithm: application of screening for aortic dissection. BMC Medical Informatics and Decision Making. 2022; 22(1):1-6.
[28] Davagdorj K, Lee JS, Pham VH, Ryu KH. A comparative analysis of machine learning methods for class imbalance in a smoking cessation intervention. Applied Sciences. 2020; 10(9):1-20.
[29] Xu Z, Shen D, Nie T, Kou Y. A hybrid sampling algorithm combining M-SMOTE and ENN based on random forest for medical imbalanced data. Journal of Biomedical Informatics. 2020; 107:103465.
[30] Johnson JM, Khoshgoftaar TM. The effects of data sampling with deep learning and highly imbalanced big data. Information Systems Frontiers. 2020; 22(5):1113-31.
[31] Shaikh S, Daudpota SM, Imran AS, Kastrati Z. Towards improved classification accuracy on highly imbalanced text dataset using deep neural language models. Applied Sciences. 2021; 11(2):1-20.
[32] Chyon FA, Suman MN, Fahim MR, Ahmmed MS. Time series analysis and predicting COVID-19 affected patients by ARIMA model using machine learning. Journal of Virological Methods. 2022; 301:1-6.
[33] Ahmed DM, Hassan MM, Mstafa RJ. A review on deep sequential models for forecasting time series data. Applied Computational Intelligence and Soft Computing. 2022; 2022:1-19.
[34] Susan S, Kumar A. The balancing trick: optimized sampling of imbalanced datasets-a brief survey of the recent state of the art. Engineering Reports. 2021; 3(4):1-24.
[35] Alharbi F, Ouarbya L, Ward JA. Comparing sampling strategies for tackling imbalanced data in human activity recognition. Sensors. 2022; 22(4):1-20.
[36] Liang XW, Jiang AP, Li T, Xue YY, Wang GT. LR-SMOTE—an improved unbalanced data set oversampling based on K-means and SVM. Knowledge-Based Systems. 2020; 196:105845.
[37] Hussein HI, Anwar SA, Ahmad MI. Imbalanced data classification using SVM based on improved simulated annealing featuring synthetic data generation and reduction. CMC-Computers Materials & Continua. 2023; 75(1):547-64.
[38] Zhao C, Shuai R, Ma L, Liu W, Wu M. Improving cervical cancer classification with imbalanced datasets combining taming transformers with T2T-ViT. Multimedia Tools and Applications. 2022; 81(17):24265-300.
[39] Christianto Y, Rusli A. Evaluating RNN architectures for handling imbalanced dataset in multi-class text classification in Bahasa Indonesia. International Journal of Advanced Trends in Computer Science and Engineering. 2020: 8418-23.
[40] Gao H, Li Y, Lu H, Zhu S. Water potability analysis and prediction. Highlights in Science, Engineering and Technology. 2022; 16:70-7.
[41] https://www.kaggle.com/datasets/adityakadiwal/water-potability. Accessed 02 March 2023.
[42] Patel J, Amipara C, Ahanger TA, Ladhva K, Gupta RK, Alsaab HO, et al. A machine learning-based water potability prediction model by using synthetic minority oversampling technique and explainable AI. Computational Intelligence and Neuroscience. 2022; 2022:1-15.
[43] Rawat N, Kazembe MD, Mishra PK. Water quality prediction using machine learning. International Journal for Research in Applied Science and Engineering Technology. 2022; 10(VI):4173-87.

[44] Wang H, Zhao Y, Zhou Y, Wang H. Prediction of urban water accumulation points and water accumulation process based on machine learning. Earth Science Informatics. 2021; 14:2317-28.
[45] Moeini M, Shojaeizadeh A, Geza M. Supervised machine learning for estimation of total suspended solids in urban watersheds. Water. 2021; 13(2):1-24.
[46] Mensi A, Tax DM, Bicego M. Detecting outliers from pairwise proximities: proximity isolation forests. Pattern Recognition. 2023; 138:109334.
[47] Buschjäger S, Honysz PJ, Morik K. Randomized outlier detection with trees. International Journal of Data Science and Analytics. 2022; 13(2):91-104.
[48] Gao R, Zhang T, Sun S, Liu Z. Research and improvement of isolation forest in detection of local anomaly points. In journal of physics: conference series 2019 (pp. 1-6). IOP Publishing.
[49] Niculescu-mizil A, Caruana R. Predicting good probabilities with supervised learning. In proceedings of the 22nd international conference on machine learning 2005 (pp. 625-32).
[50] Mulugeta G, Zewotir T, Tegegne AS, Juhar LH, Muleta MB. Classification of imbalanced data using machine learning algorithms to predict the risk of renal graft failures in Ethiopia. BMC Medical Informatics and Decision Making. 2023; 23(1):1-7.
[51] Alipour A, Ahmadalipour A, Abbaszadeh P, Moradkhani H. Leveraging machine learning for predicting flash flood damage in the Southeast US. Environmental Research Letters. 2020; 15(2):1-12.
[52] Ahmed U, Mumtaz R, Anwar H, Shah AA, Irfan R, García-nieto J. Efficient water quality prediction using supervised machine learning. Water. 2019; 11(11):1-14.
[53] Gakii C, Jepkoech J. A classification model for water quality analysis using decision tree. Euro Journal of Computer Science and Information Technology. 2019; 7(3):1-8.
[54] Jaloree S, Rajput A, Gour S. Decision tree approach to build a model for water quality. Binary Journal of Data Mining & Networking. 2014; 4(1):25-8.
[55] Khan TM, Xu S, Khan ZG. Implementing multilabeling, ADASYN, and relieff techniques for classification of breast cancer diagnostic through machine learning: efficient computer-aided diagnostic system. Journal of Healthcare Engineering. 2021; 2021:1-15.
[56] Peng CY, Park YJ. A new hybrid under-sampling approach to imbalanced classification problems. Applied Artificial Intelligence. 2022; 36(1):1-18.
[57] Zhang A, Yu H, Huan Z, Yang X, Zheng S, Gao S. SMOTE-RkNN: a hybrid re-sampling method based on SMOTE and reverse K-nearest neighbors. Information Sciences. 2022; 595:70-88.
[58] Pan T, Zhao J, Wu W, Yang J. Learning imbalanced datasets based on SMOTE and Gaussian distribution. Information Sciences. 2020; 512:1214-33.
[59] Wang K, Tian J, Zheng C, Yang H, Ren J, Li C, et al. Improving risk identification of adverse outcomes in chronic heart failure using SMOTE+ ENN and machine learning. Risk management and Healthcare Policy. 2021:2453-63.
[60] Branco P, Torgo L, Ribeiro RP. A survey of predictive modeling on imbalanced domains. ACM Computing Surveys (CSUR). 2016; 49(2):1-50.
[61] Silva FT, Song H, Perello-nieto M, Santos-rodriguez R, Kull M, Flach P. Classifier calibration: a survey on how to assess and improve predicted class probabilities. Machine Learning. 2023:1-50.
[62] Alqarni AA, Yadav OP, Rathore AP. Application of isotonic regression in predicting corrosion depth of the oil refinery pipelines. In annual reliability and maintainability symposium 2022 (pp. 1-6). IEEE.
[63] Mahmudah KR, Indriani F, Takemori-sakai Y, Iwata Y, Wada T, Satou K. Classification of imbalanced data represented as binary features. Applied Sciences. 2021; 11(17):1-13.
[64] Wegier W, Ksieniewicz P. Application of imbalanced data classification quality metrics as weighting methods of the ensemble data stream classification algorithms. Entropy. 2020; 22(8):1-17.
[65] Ri J, Kim H. G-mean based extreme learning machine for imbalance learning. Digital Signal Processing. 2020; 98:102637.
[66] Aridas CK, Karlos S, Kanas VG, Fazakis N, Kotsiantis SB. Uncertainty based under-sampling for learning naive bayes classifiers under imbalanced data sets. IEEE Access. 2019; 8:2122-33.
[67] Ala'raj M, Abbod MF, Majdalawieh M. Modelling customers credit card behaviour using bidirectional LSTM neural networks. Journal of Big Data. 2021; 8(1):1-27.
[68] Rožanec JM, Bizjak L, Trajkova E, Zajec P, Keizer J, Fortuna B, et al. Active learning and novel model calibration measurements for automated visual inspection in manufacturing. Journal of Intelligent Manufacturing. 2023:1-22.

C. Kaleeswari is currently pursuing PhD at the Department of Computational Logistics, Alagappa University, Tamilnadu, India. Her areas of research interest are Data Science, Machine Learning, Internet of Things, Environmental Assessment Modelling, and Information Security.
Email: kalees94chinna@gmail.com

K. Kuppusamy received the PhD degree in computer science and engineering from Alagappa University, Tamilnadu, India in 2007. He has 34 years of Teaching and 20 years of research experience. From 2007 to 2022 he worked with various positions in Alagappa University as a Professor, Director of Computer Centre, Chairperson of Computer Sciences. His areas of interest are Cloud Security, Algorithms, Information and Network Security, and Software Engineering.
Email: kkdiksamy@yahoo.com

A. Senthilrajan received the PhD degree in computer science and engineering from Alagappa University, Tamilnadu, India in 2011. He has 25 years of teaching and 10 years of research experience. He is currently working as a Professor and Head in the Department of Computational Logistics and Director of Management Information System (MIS) in Alagappa University, Tamilnadu, India. His areas of interest are Image Processing, Networks, Artificial Intelligence.
Email: agni_senthil@yahoo.com

Appendix I

S. No.	Abbreviation	Description
1	AdaSyn	Adaptive Synthetic Sampling
2	ANN	Artificial Neural Networks
3	AUC	Area Under the Curve
4	BRA	Balanced Random Accuracy
5	BS	Brier Score
6	CCG	Cervical Cell Generation
7	CNN	Condensed Nearest Neighbour
8	DL	Deep Learning
9	DT	Decision Tree
10	EC	Electrical Conductivity
11	ECE	Expected Calibration Error
12	ENN	Edited Nearest Neighbour
13	GBT	Gradient Boosting Tree
14	GNB	Gaussian Naive Bayes
15	GRU	Gated Recurrent Unit
16	G-mean	Geometric Mean
17	IF	Isolation Forest
18	IR	Isotonic Regression
19	KNN	K-Nearest Neighbour
20	LR	Logistic Regression
21	LSTM	Long Short Term Memory
22	MCE	Maximum Calibration Error
23	ML	Machine Learning
24	MLP	Multi-Layer Perceptron
25	MSE	Mean Squared Error
26	NB	Naive Bayes
27	NN	Nearest Neighbour
28	NOM	Natural Organic Matter
29	NP	Near Pseudo
30	PH	Potential Hydrogen
31	PS	Platt Scaling
32	PSO	Particle Swarm Optimization
33	RF	Random Forest
34	RNN	Recurrent Neural Networks
35	ROS	Random Over Sampling
36	RUS	Random Under Sampling
37	SA	Simulated Annealing
38	SMOTE	Synthetic Minority Oversampling Technique
39	SMOTEENN	SMOTE with Edited Nearest Neighbour
40	SVC	Support Vector Classifier
41	SVM	Support Vector Machine
42	TDS	Total Dissolved Solids
43	THMs	Trihalomethanes
44	TOC	Total Organic Carbon
45	Tomek	T-link
46	WHO	World Health Organisation