IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 April 2024, accepted 9 April 2024, date of publication 22 April 2024, date of current version 29 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3391749

RESEARCH ARTICLE

Ok-NB: An Enhanced OPTICS and k-Naive Bayes Classifier for Imbalance Classification With Overlapping

ZAHID AHMED[1], BIJU ISSAC[2], (Senior Member, IEEE), AND SUFAL DAS[1]
[1]Department of Information Technology, North-Eastern Hill University, Shillong 793022, India
[2]Department of Computer and Information Sciences, Northumbria University, NE1 8ST Newcastle upon Tyne, U.K.

Corresponding author: Sufal Das (sufal.das@gmail.com)

ABSTRACT Class imbalance problems have received a lot of attention throughout the last few years. It poses considerable hurdles to conventional classifiers, especially when combined with overlapping instances, where the complexity of the classification task increases. In this study, we have proposed a novel density-based method that combines the Ordering Points To Identify the Clustering Structure (OPTICS) algorithm with the Naive-Bayes approach to effectively handle overlapped and imbalanced problems at the same time, known as OPTICS-based k-Naive Bayes (Ok-NB). The Ok-NB method is used to correctly identify and construct the training data into overlapping and non-overlapping groups based on their density and reachability, while the Naive-Bayes technique is used to correctly map the test data samples to the appropriate class for accurate output. It offers adaptability and reliability in classifying complex datasets with overlapping and imbalanced properties. Cluster-based proximity assessment and probabilistic classification are combined to improve classification accuracy and guarantee that the most reliable neighbours' opinions are given the greatest weight during the decision-making process. Extensive experiments were conducted on 21 benchmark datasets and the experiment results demonstrate how effectively the suggested approach works to achieve high classification accuracy. This proves the effectiveness and superiority of this proposed approach compared to existing state-of-the-art methods in tackling overlapping and imbalance challenges in classification tasks.

INDEX TERMS Classification, imbalanced data, overlapped data, OPTICS, Naive-Bayes.

I. INTRODUCTION

Class imbalance in the dataset is a prevalent and challenging issue in machine learning, where the instances of the classes are not evenly distributed [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. Most of the time, traditional machine learning algorithms are more attracted to the majority class (classes having a large number of instances [10], [12]), and the minority class (classes having a few numbers of instances [10], [12]) is ignored which results in reduced accuracy and performance [7], [8], [13]. This problem is encountered in a wide range of real-world applications like medical diagnosis [14], [15], fraud detection [16], [17], fault prediction [18], [19], text classification [20], [21] etc. To diminish the impact of class imbalance, various techniques have been proposed which include data-level, algorithm-level, and hybrid approaches. The data-level approaches deal with data preprocessing, such as oversampling and undersampling, to balance the class distribution. Although it is effective in some scenarios, it may suffer from overfitting problems, increased computational complexity, or loss of information [4], [7], [8], [10], [13], [22], [23]. In contrast, algorithm-level approaches directly address the class imbalance within the learning algorithm itself. These approaches modify the algorithms' learning mechanisms to adapt the imbalanced data and improve performance across all classes. By enhancing the base learning algorithms, algorithm-level techniques seek to achieve a more balanced

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

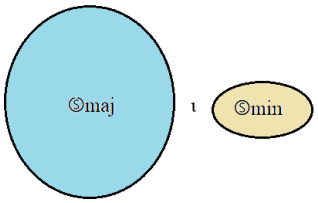

FIGURE 1. Noisy instance.

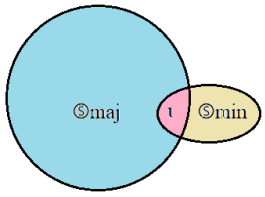

FIGURE 2. Overlapped instance.

and accurate classification for both majority and minority classes without compromising generalization capabilities [4], [7], [8], [10], [13], [22], [23]. A hybrid approach combines these two approaches to tackle this issue [4], [7], [8], [10], [13], [22], [23]. While all the above-mentioned approaches aim to address the challenges posed by imbalanced data, the algorithm-level approach offers distinct advantages that make it more effective and practical in various real-world scenarios [24], [25]. The algorithm-level approach directly alters the learning processes of the algorithm to adapt imbalanced data, ensuring that the model is naturally capable of handling class imbalance without the need for extensive data preprocessing techniques. In contrast, data-level approaches like oversampling or undersampling might cause data duplication or loss, which could result in overfitting or important data deterioration [12], [24], [25]. By specifically taking into account class imbalance throughout the learning phase, the algorithm-level approach may more evenly distribute predictive performance across all classes, enhancing accuracy and recall for both majority and minority classes [12], [24], [25]. In the data-level approach, the training data set may need to be altered, or synthetic data may need to be created, which can greatly increase training time and computer resources. Instead of requiring considerable data modification, the algorithm-level method makes use of the learning algorithm's built-in features, which makes this approach more effective [12].

Prediction accuracy is substantially hampered by the presence of noise and overlapped instances in the data set. An instance is considered noise if it does not belong to both the minority class and the majority class and an instance is considered an overlapped instance if it belongs to both the majority and minority class [4], [7], [8], [13], [22], [23]. If $Ⓢ_{maj}$ and $Ⓢ_{min}$ are the two set of majority and minority instance respectively then instance $ι$ is considered noise if:

$$ι \notin Ⓢ_{maj} \wedge Ⓢ_{min} \tag{1}$$

An instance $ι$ is considered an overlapped instance if:

$$Ⓢ_{maj} \cap Ⓢ_{min} = ι \tag{2}$$

Figure 1 represents noise instance, whereas Figure 2 represents overlapped instance. Noise magnifies prediction errors and decreases model reliability, and overlapping blurs class boundaries that lead to misclassification [26], [27], [28], [29]. The majority of research concentrates on finding new ways to solve overlapping and imbalance problems individually. However, the merging of overlapping and imbalance concerns in real-world applications makes the categorization task extremely difficult. Even though some attempts have been made to handle both issues simultaneously, implementation is not feasible due to the complexity of the algorithm structures [30], [31].

A. MOTIVATION

The proposed approach is initiated to provide a general framework to deal with the complex problems caused by imbalanced datasets with overlapping instances. Traditional classification approaches frequently fail to categorize instances properly because of the inherent complications associated with class imbalance and overlapping instances. To efficiently tackle overlapped and imbalanced situations at the same time, a density-based strategy called OPTICS-based k-Naive Bayes (Ok-NB) is proposed that combines two key components to address this problem. (i) Cluster-based preprocessing using the Ordering Points To Identify the Clustering Structure (OPTICS) [32] algorithm with a modified reachability distance function, and (ii) A probabilistic classification method based on Naive Bayes [33] approach incorporating a new weighted score concept along with the consideration of the top k weight concept.

OPTICS is a density-based clustering algorithm that can handle overlapping data in an imbalanced dataset [32]. It is designed to locate clusters depending on the number of data points present. As a result, even if the clusters are overlapped, it can still capture regions of extremely dense data, which makes it particularly helpful for handling overlapping data. It works effectively in scenarios where it may be difficult for typical distance-based algorithms to distinguish between clusters [32]. In contrast to DBSCAN (a density-based algorithm for discovering clusters in large spatial databases with noise) [34], it does not rely on a fixed neighbourhood radius. As an alternative, it employs an adjustable reachability distance that enables it to capture clusters of all sizes and forms, including overlapping clusters.

Through the process of clustering, the imbalanced and overlapped training dataset is partitioned into discrete clusters, each of which represents a different subset of instances, such as majority and minority noise, overlapped majority and minority instances, majority instances (without overlap), and minority instances (without overlap). This process enables the proposed approach to properly distinguish between noisy

data and overlapping instances, which provides a way for a more precise categorization.

To perform the subsequent task, the preprocessed clustered data is used as input for the Naive Bayes algorithm. The Naive Bayes technique is used to correctly map the test data samples to the right group for accurate output. Naive Bayes is included in the OPTICS framework because of its probabilistic foundation, which makes it possible to describe complex relationships in the dataset. Naive Bayes is especially well-suited for situations where features may exhibit varying degrees of dependence. The proposed approach leverages the advantages of probabilistic reasoning to improve the precision of cluster assignments. Users can comprehend the concepts behind classification judgments by using the probability and weighted scores provided for each class, which enhances interpretability. To customize the approach for various datasets and settings, users can modify the value of the parameter k, which determines how many top-weighted clusters are taken into account during the classification process.

The significance of the proposed strategy is the systematic methodology that allows it to manage imbalanced datasets with overlapping classes successfully. It offers a strong framework for precise classification even in difficult situations with class imbalance and overlap by combining cluster-based preprocessing and probabilistic classification procedures. This method improves the overall reliability of the classification process by increasing the accuracy of the results and provides insights into ambiguous situations.

B. CONTRIBUTIONS OF THE STUDY

The primary contributions of this study are:

- Development of an approach that enhances data clustering by considering overlapping data in an imbalanced dataset.
- Creating a strategy for effective classification.
- Development of an algorithm-level (two-step) approach to deal with the overlapped and imbalanced problem at the same time.

C. PAPER ORGANIZATION

This paper has been organized as follows: In Section II, the relevant works in this area are discussed. The details of the background study are included in section III. Section IV provides a brief overview of this recommended strategy. Extensive experiments on various datasets are presented in Section V. The results and analysis are presented in Section VI. Finally, Section VII presents the conclusion and future works.

II. RELATED WORKS

Numerous machine learning and data mining systems struggle with the issue of imbalanced data. When classes are imbalanced in a dataset, it can lead to biased model performance, where the minority class may be misclassified. As mentioned in the previous section, there have been many different strategies suggested to address this issue. The data-level and algorithm-level approaches are the most widely used techniques [35].

The data-level approach concentrates on altering the dataset itself to adjust class distributions before supplying the dataset to a classifier [10], [12]. Making synthetic samples for the minority class, the Synthetic Minority Over-sampling Technique (SMOTE) [36] is a pioneering approach to solving the issue of class imbalance. By synthesizing minority class samples, SMOTE provides an effective solution for class imbalance; however, its efficacy depends on robustness to high-dimensional feature spaces and proper parameter selection [36]. SMOTE has undergone numerous additions and modifications throughout time to improve its functionality and suitability for a range of situations. These variations include Borderline-SMOTE (B-SMOTE) [37], Kernel-based SMOTE (K-SMOTE) [38], Support Vector Machine -SMOTE (SVM-SMOTE) [39], Adaptive Synthetic Sampling Approach (ADASYN) [40] etc. These variations offer distinct advantages and disadvantages. Borderline-SMOTE identifies those instances of minority class close to the decision boundary, where the classification task is more difficult. Then it creates synthetic instances for these borderline circumstances. Although it decreased noise and enhanced classification performance, its efficacy can differ based on parameter settings and dataset properties [37]. By creating artificial minority class instances in a high-dimensional feature space specified by a kernel function, k-SMOTE enables the construction of sophisticated and non-linear decision boundaries. Although it offers enhanced classification performance and non-linearity, it exhibits difficulties with computing complexity, kernel selection, and interpretability [38]. SVM-SMOTE combines the SMOTE technique with SVM classification to create synthetic instances that are more informative for the classifier. It may enhance classification performance by creating synthetic instances in areas where the SVM classifier has low confidence. While it combines the strength of SMOTE and SVM, its effectiveness in some scenarios may be limited by its sensitivity to classifier parameters and computational cost [39]. ADASYN is considered a prominent approach to dealing with the problem of class imbalance. It concentrated on instances in the minority class where accurate classification is more challenging. It creates synthetic instances for minority classes where the class distribution is most imbalanced. It has difficulties with computational complexity, noise sensitivity, and parameter selection, but it also provides flexibility and attention to difficult instances [40].

The algorithm-level approaches can change the learning process and provide more accurate, impartial, and fair models. Existing approaches like k-Nearest Neighbor (k-NN) [41], Support Vector Machine (SVM) [42], and Random Forest [43] have significantly inspired the creation of new algorithms and strategies to address this problem.

These algorithms have not only made way for cutting-edge methods but have also directly advanced the processing of imbalanced data.

In typical machine learning problems, the k-NN is one of the most effective and straightforward classifiers but the performance of k-NN suffers a lot if the data are imbalanced. To solve this issue, the K Exemplar-based Nearest Neighbor algorithm (ENN) [51] is proposed. As a pattern-oriented strategy, it is characterized as relying on amplifying the influence of minority class samples. The pivot minority class instances are chosen and their boundaries are expanded into Gaussian balls as part of the approach's operation. Another pattern-oriented approach that is comparable to ENN is the Positive-biased Nearest Neighbor (PNN) [52]. However, it does not include a training step. Compared to ENN, PNN is a faster approach. The distribution-oriented methods, which rely on collecting meaningful prior knowledge of the data distribution, stand in contrast to the pattern-oriented methods. One of these techniques is the Class-Based Weighted k Nearest Neighbor, which balances the instances based on the estimated k-NN misclassification rate [53]. Other examples of distribution-oriented approaches include Class Conditional Nearest Neighbor Distribution (CCNND) [54] and Informative k Nearest Neighbor-localized version (Ll-kNN) [55]. k Rare Class Nearest Neighbor Classification (K-RNN) [44] is one approach that focuses on locating and classifying uncommon occurrences in an imbalanced dataset. It seeks to enhance the classification of minority groups by taking into account each data point's k-nearest neighbours, but due to the computationally demanding nature of k-nearest neighbour identification, it might have scalability problems when working with huge datasets. Additionally, it might not function effectively when there is a strong class imbalance since the rare class may not be sufficiently isolated from the majority class in feature space. In order to choose a balanced training set from imbalanced data and enhance model performance, a Memetic Approach for Training Set Selection in Imbalanced Data Sets (BQI-GSA) [46] was proposed that combines genetic algorithms and simulated annealing but the success of BQI-GSA depends on parameter settings, which can be dataset-specific and computationally costly. In order to specify a constant radius for closest neighbour classification in imbalanced datasets, Gravitational Fixed Radius Nearest Neighbor for Imbalanced Problem (GFRNN) [47] uses the idea of gravity forces to increase the overall accuracy of imbalanced data classification jobs by adjusting the neighbourhood size for each data point based on its class distribution, but it has trouble to handle regions with different densities or high-dimensional data.

In order to improve the categorization of minority classes, the Neighbors Progressive Competition Algorithm (NPC) [48] gradually competes with neighbouring samples to address the imbalanced data. This strategy encourages the repeated adjustment of class borders and has demonstrated potential for enhancing classification performance on imbalanced data, but it is also sensitive to the initial neighbourhood size selection, which may affect the overall performance. Least Squares KNN-Based Weighted Multi-class Twin SVM (LS-KWMTSVM) [49] combines the ideas of K-nearest neighbours and twin SVMs and uses a least squares method to apply various weights to the classes but it is not suitable for large datasets. Density-Based Adaptive K Nearest Neighbor (DBANN) [50] can handle overlapping problems in imbalanced datasets by modifying KNN to capture overlapped regions. In terms of high-dimensional data or extensive overlaps, it may create difficulties.

SVM is another effective classifier that looks for the best way to classify data points into distinct groups. One of the most significant weaknesses of SVM is its tendency to favour the majority class when handling unbalanced datasets. A number of approaches have been proposed based on SVM such as Fuzzy Support Vector Machines (F-SVM) [56], Fuzzy total Margin based Support Vector Machine (FM-SVM) [57], Entropy-based Fuzzy Least Squares Twin Support Vector Machine (EFLT-SVM) [58] etc. The F-SVM algorithm is recommended primarily for handling noise and outliers. Different fuzzy membership values are assigned to the instances in this case to characterize their significance. The performance of this method depends on fuzzy membership value calculation techniques. Outliers and noise typically have lower fuzzy membership values than the other samples. It also struggles with issues related to imbalanced data. Support Vector Machine-Based Optimized Decision Threshold Adjustment Strategy (SVM-OTHR) [45] improves the efficiency of support vector machines (SVM) in datasets with imbalances by optimizing the decision threshold. By tweaking the decision boundary, this adjustment approach enables the SVM classifier to manage class imbalances more effectively, but the kernel function and regularisation parameters need to be carefully chosen. During parameter adjustment, class imbalance may require special handling. An overview of all the algorithmic-level approaches explained above is summarized in table 1.

It can be observed from table 1 that each of the mentioned approaches has made a substantial contribution to this domain. There are still certain research gaps. The sensitivity of these approaches towards parameter settings and the requirement for robustness across different datasets and complexity is one notable difference between them. For instance, the definition of rare classes and the value of k can affect the efficacy of K-RNN. The performance of SVM-OTHR is dependent on the choice of kernel and parameter adjustments. Likewise, approaches like BQI-GSA and GFRNN face difficulties concerning convergence and sensitivity to dataset properties, which restricts their use in many problem domains. Furthermore, the complexity of LS-KWMTSVM's parameter tuning procedure and the difficulty of NPC's conflict mechanism design emphasize the need for more flexible and scalable approaches. Though current approaches provide useful solutions for class imbalance,

TABLE 1. Literature summery Table.

Approach	Findings	Limitations	Remarks
K-RNN [44]	Potentially promising strategy with good performance for handling rare classes.	High computational complexity, sensitivity to k, and rare class definition. Overlapping problems not addressed.	Not addressed overlapping problems.
SVM-OTHR [45]	Adaptable threshold.	Parameter tuning is required and not suitable for overlap data.	Overlapping issues are not taken into consideration.
BQI-GSA [46]	Improves classification performance by choosing representative instances.	Complexity of parameter tuning, possible problems with convergence, and lack of attention to overlap.	Insufficient focus on the overlap.
GGFRNN [47]	Robust to noise reduction and class imbalance.	Computational complexity, sensitivity to parameters and dataset characteristics, and Overlapping issues ignored.	Ignored overlapping challenges.
NPC [48]	Adaptive to the features of the local dataset.	Sensitivity to parameters, complexity in the competitive mechanisms, and not targeted overlap challenges.	Overlapping issues are not taken into consideration.
LS-KWMTSVM [49]	Incorporates flexibility in multiclass imbalance and local and global information.	Complexity in parameter tuning, potential sensitivity to datasets, and minimal overlap focused.	Insufficient focus on the overlap.
DBANN [50]	Can utilize global density information and flexible to data distributions.	Parameter sensitivity, computational complexity, and overlapping problems were ignored.	Overlapping challenges sidelined.
k-NN [41]	Simple, non-parametric, and versatile	Sensitivity to k selection, computationally complex, and overlapping issue is not targeted.	Ignored overlapping challenges.
SVM [42]	Effective in high-dimensional space.	complexity in parameter tuning and sensitivity to imbalance.	Insufficient focus on the overlap.
RF [43]	Can manage high-dimensional data, resistant to overfitting, and can represent intricate relationships.	Not appropriate for imbalance or overlapping data, prone to overfitting on noisy data.	Overlapping challenges ignored.

they frequently ignore the complex nature of overlapping instances.

III. BACKGROUND

A. OPTICS

A density-based method called OPTICS (Ordering Points To Identify the Clustering Structure) [32] is a dominant density-based clustering technique used in data mining and machine learning. It has the potential to recognize noisy instances and clusters of different sizes and shapes, including clusters with irregular structures. It is an extension of the popular Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [34] technique. While DBSCAN is quite good at locating dense clusters, it provides a more thorough picture of the clustering structure by generating a reachability plot, which includes useful information about the density distribution of the data set. The MinPts and epsilon ε radius are the parameters used by this algorithm, where ε describe the area around each point, and MinPts is the minimal number of points necessary to construct a dense region to determine how close points must be to one another in order to associate with the same cluster [32]. The fundamental ideas behind OPTICS are:

1) CORE DISTANCES

The local density surrounding a data point ι is measured by its core distance. It is defined as the distance between ι and its MinPts, where MinPts represents the lowest number of data points needed to make a dense zone and is a user-defined parameter. If a data point is a core point, it can be determined using the core distance [32].

$$\text{core distance}(\iota) = \text{distance}(\iota, N_{\text{MinPts}(\iota)}) \tag{3}$$

where, $N_{\text{MinPts}(\iota)}$ is the collection of MinPts that are closest to ι.

2) REACHABILITY DISTANCE (RD)

It calculates the density of the relationship between two data points. A denser area is indicated by a shorter reachability distance [32].

$$\text{RD}(\iota_i, \iota_j) = \max(\text{core distance}(\iota_j), \text{distance}(\iota_i, \iota_j)) \tag{4}$$

where, $\text{distance}(\iota_i, \iota_j)$ = The Euclidean distance or another distance metric between the data points ι_i and ι_j.

B. NAIVE BAYES CLASSIFIER

Bayesian statistics and probabilistic reasoning serve as the foundation of the Naive Bayes method. It uses the Bayes theorem and the "naive" assumption of feature independence to determine the conditional probability of a class given by the data. This enables it to classify objects or make predictions based on the most likely class [33]. The fundamental job of this classifier can be considered as follows:

1) DATA PREPARATION

Preprocess the dataset, ensuring features are independent and identically distributed. Transform categorical variables into numerical representations.

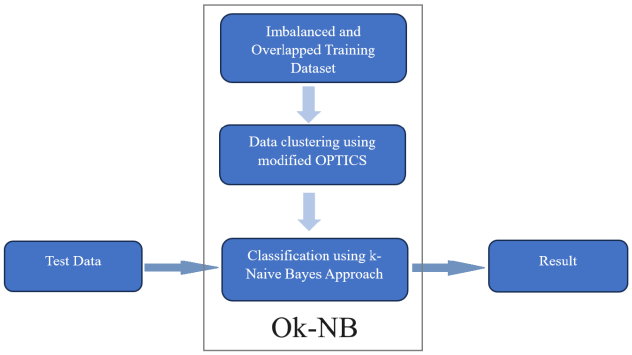

FIGURE 3. Block diagram of the proposed Ok-NB method.

2) PROBABILITY ESTIMATION

Calculate the class probabilities (prior probabilities) and the conditional probabilities for each feature given the class. This involves estimating the probability distributions, such as Gaussian for continuous data or multinomial for discrete data.

As this algorithm is based on Bayes' theorem, therefore based on predictor (prior probability) and observable data (conditional probability), it determines the likelihood of a hypothesis (class) [33].

$$P(\zeta|\varkappa) = \frac{[P(\varkappa|\zeta) \cdot P(\zeta)]}{P(\varkappa)} \tag{5}$$

where, $P(\zeta|\varkappa)$ = Posterior probability of a class ζ given predictor $\varkappa$.

$P(\varkappa|\zeta)$ = Likelihood of predictor $\varkappa$ of a given class ζ.

$P(\zeta)$ = Prior probability of class ζ.

$P(\varkappa)$ = Total probability of predictor $\varkappa$ (Normalization factor).

To determine how likely a specific feature or attribute value is given a class, conditional probabilities are used.

$$P(\zeta|\varkappa) = \frac{\text{Number of instances with feature } \varkappa \text{ and class } \zeta}{\text{Total number of instances in class } \zeta} \tag{6}$$

The likelihood of a class occurring without taking into account any particular attribute is represented by the prior probability.'

$$P(\zeta) = \frac{\text{Number of instances in class } \zeta}{\text{Total number of instances}} \tag{7}$$

With dependent feature vectors $\varkappa_1$ through $\varkappa_n$ and class variable, the probability $\forall$ i is:

$$P(\zeta|\varkappa_1, \ldots, \varkappa_n) = \frac{P(\zeta)\Pi_{i=1}^{n} P(\varkappa_i|\zeta)}{P(\varkappa_1, \ldots, \varkappa_n)} \tag{8}$$

These probabilities are determined by Naive Bayes for each class, and the class with the highest posterior probability is designated as the predicted class for a specific instance. When doing a classification task, this procedure is repeated for each class.

IV. PROPOSED METHOD

Previous studies have clearly demonstrated the importance of query neighbors in k-nearest neighbors (k-NN) [41] classification. These studies have repeatedly highlighted that nearby data points play a vital role in the k-NN base algorithm's decision-making process. For reliable and accurate classification results in k-NN, it is crucial to understand and utilize the connections and properties of surrounding data points [52], [59]. To deal with overlapping data, the Adaptive k-Nearest Neighbours (A-kNN) [60] approach dynamically changes the distance metric based on how much the data points overlap and also calculates and relies on a reliable coefficient (r_i) for each training instance (ι_i). It is the distance from ι_i to its closest neighbour ι_j that belongs to a different class.

$$r_i = \text{min\_distance}(\iota_i, \iota_j) \tag{9}$$

Some studies also focused on dynamically modifying the query neighbours based on the precise level of class imbalance and data overlap in the dataset by considering r_i and the majority vote $f(\iota)$ for each training instance (ι_i) for classification [50]. This concept has inspired the addition of a weighted score to the suggested approach, which will be determined by the following equation.

$$W(\iota) = P(\zeta) \cdot P(\varkappa|\zeta) \tag{10}$$

This adaptability guarantees that the neighbourhood selection can change in accordance with the distinctive properties of the data, ultimately producing more accurate and reliable classification results in situations where conventional k-nearest neighbour approaches may fail.

Motivated by this, the proposed OPTICS-based k-Naiv Base (Ok-NB) approach has been designed that combines the power of OPTICS [32] for cluster-based proximity assessment and pairs with the probabilistic classification capabilities of Naive Bayes [33] to identify and rank those nearby data points that are most trustworthy and instructive for making classification decisions. Naive Bayes is renowned for its simplicity and computational efficiency. It can also handle datasets with an extensive feature count that helps to overcome issues with high-dimensional datasets.

To address imbalanced and overlapping situations simultaneously, the r_i concept is adopted and utilized to modify the reachability distance described in equation 4 of the OPTICS algorithm and consider the equation 11 for overlapped reachability distance (ORD) of this proposed approach.

$$\text{ORD}(\gamma) = \frac{\max(\text{core distance}(\iota_j), \text{distance}(\iota_i, \iota_j))}{r_i} \tag{11}$$

With the new reachability distance, it is now possible to accomplish finer-grained clustering, better separation of overlapping clusters, better noise treatment, adjustable control over reliability, effective handling of imbalanced data, robustness to outliers, and increased cluster quality. These benefits make it a useful method for grouping complex datasets with overlapping and imbalanced properties.

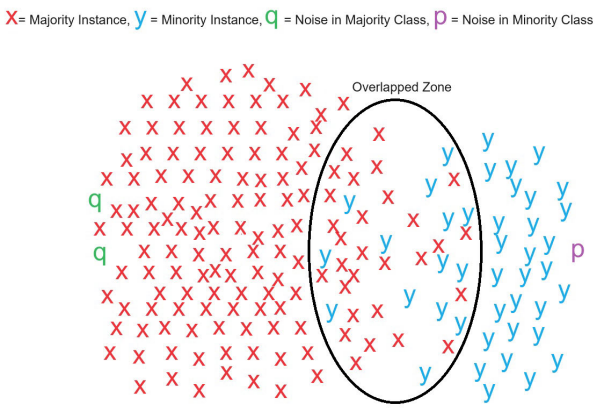

FIGURE 4. Imbalanced dataset with overlapped and noise.

The integration of Naive Bayes into the modified OPTICS framework provides the capability to effectively represent feature relationships that provide a scalable resolution for sophisticated datasets. It also offers a sound foundation for improving the precision of cluster assignments and strengthening the interpretability of the outcomes. To improve the overall accuracy and reliability of the classification process, the opinions of the most trustworthy neighbours are given the greatest weight and the final classification is done by considering the maximum k weight. This comprehensive approach guarantees solid and dependable final classification outcomes, even in difficult data situations.

This proposed method is a two-step process that consists of the following phases:

1) Data Clustering using Modified OPTICS Technique.
2) Classification using Proposed k-Naive Bayes Approach.

Figure 3 shows the block diagram for the suggested approach.

A. DATA CLUSTERING USING MODIFIED OPTICS TECHNIQUE

In the first step, a binary imbalanced dataset with overlapping and noisy instances is divided into different clusters using the OPTICS algorithm with a modified reachability distance, where the reachability distance is considered as mentioned in equation 11. With this modification, it is possible to take into account the accuracy of reachability when locating clusters in an imbalanced dataset with overlapping data. These categories are intended to capture various features of the dataset, such as minority and majority instances as well as overlapping, and noisy instances. Accordingly, training data can be clustered up to six clusters. (a) Majority noise, (b) Minority noise, (c) Overlapped majority instances, (d) Overlapped minority instances, (e) Majority instances (without overlapped), and (f) Minority instances (without overlapped). Figure 4, demonstrates an imbalanced data set with overlapped and noise.

Grouping the data into various clusters can reduce the overlap and improve the ability to distinguish the classes more precisely. It offers a detailed illustration of the dataset. Because of its flexibility, the algorithm can successfully detect differences between overlaps and imbalances and modify its approach as necessary. While some clusters might place more emphasis on areas with less overlap, certain clusters might concentrate on areas with considerable overlap. With distinct clusters, it can customize its classification strategy to the traits of each cluster, enabling more context-aware predictions. It will also help in making more accurate and trustworthy predictions. This method can help to identify instances of the minority class more precisely and reduce misclassification.

Consider a two-dimensional dataset with data points A, B, C, D, and E forming two clusters, X and Y. B belongs to both clusters, but D and E do not belong to any clusters. Now to recognize the overlaps and noise, please note the following:

- To measure the density of each point at first, the core distances of each instance must be calculated.
- The reliable coefficient and the reachability distances are calculated by taking into account both actual and core distances.
- Reachability distances determine the density. Dense places that have relatively low reachability distances are considered clusters.
- The distance between an instance of one class and an instance of another class is represented by the reliable coefficient. Points with low reliable coefficients with several clusters are frequently noticed as overlapping instances, indicating that these examples are dense in multiple clusters.
- Lower density is indicated by points with higher reachability distances representing noise.

The reachability distance and reliable coefficient of B are relatively small in both clusters (indicating that B is a dense point for both X and Y), and if a point is dense in several clusters, it results in an overlap. D and E have high reachability distances with their neighbors and are recognized as noise.

The algorithm for clustering the instances is mentioned in the algorithm 1.

In the initial stage, the imbalanced dataset is loaded for the subsequent clustering task. After that, the OPTICS algorithm is applied to process the dataset using a modified reachability distance stated in the equation 11. This modification improves the clustering capabilities of this approach, especially for datasets with a high imbalance ratio. The threshold values for cluster density are also defined in this phase. To produce specified clusters, parameters like γ (maximum distance between two instances to be considered neighbours) and MinPts (minimum number of points to constitute a cluster) are used. Depending on the unique properties of the dataset, these thresholds can be modified. The lists for each category are then initialized after that. In order to group instances into six aforementioned clusters, six empty clusters are created. For the purpose of storing

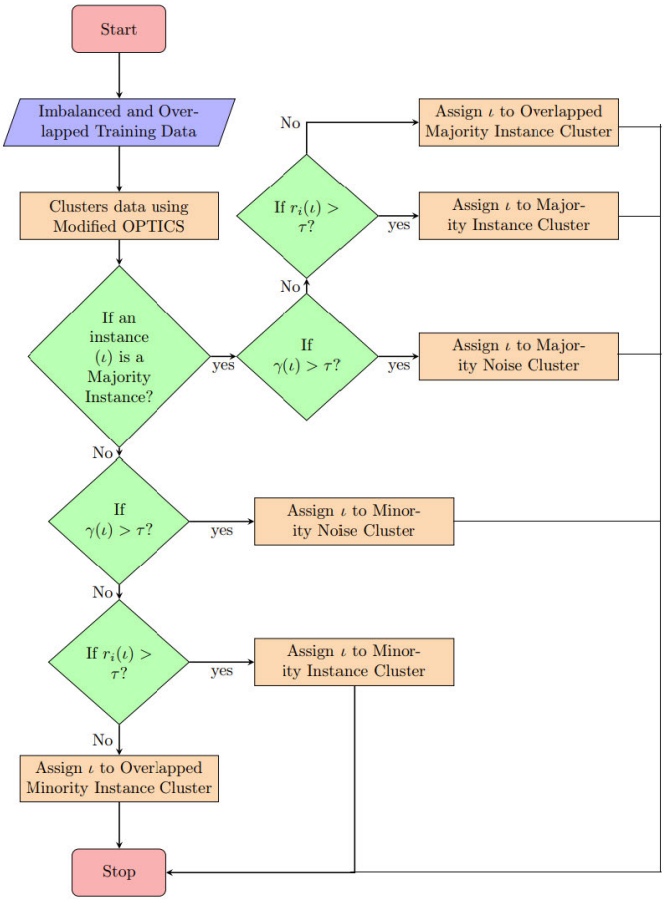

FIGURE 5. Flowchart of the proposed data clustering approach.

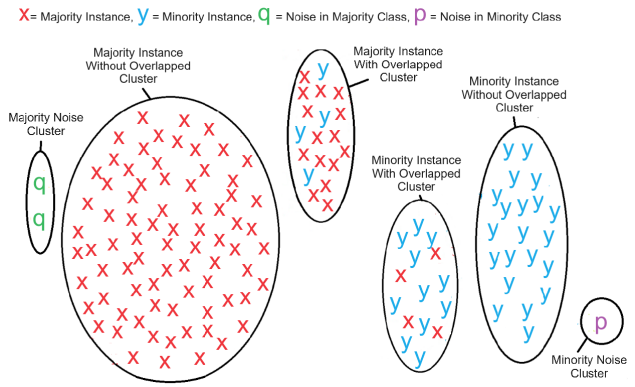

FIGURE 6. Clustering instances with the modified OPTICS algorithm.

instances and calculating the mean for each cluster, lists are made. To keep track of the processed data, a blank set is initialized. The reachability distance γ is determined for each instance in the dataset by using equation 11. Based on the separation of the instances, the reliable coefficient r_i is used to calculate the distance between the instances by using equation 9. This phase modifies the distance calculation to take the γ into account for the reliability of each instance. For storing instances for further processing, a blank queue is created. The clusters are eventually formed by using the equations 3 and 11 and based on the following conditions.

- Clusters with majority class instances with higher reachability distances γ are considered as majority noise.
- Clusters with minority class instances with higher reachability distances γ are considered as minority noise.
- Clusters with majority class instances, low reachability distances γ, and low reliable coefficient (r_i) are considered as overlapped majority instances.
- Clusters with minority class instances, low reachability distances γ, and low reliable coefficient (r_i) are considered as overlapped minority instances.
- Clusters with majority class instances, low reachability distances γ, and high reliable coefficient (r_i) are considered as majority instances.
- Clusters with minority class instances, low reachability distances γ, and low reliable coefficient (r_i) are considered as minority instances.

Finally, the aforementioned six clusters will be created.

Algorithm 1 Data Clustering Using Modified-OPTICS

Input:
Dataset $\mathbb{D}$, where $\iota \in \mathbb{D}$ (ι is an instance) with class labels ι_{maj} (majority instance) and ι_{min} (minority instance).
MinPts (Minimum number of points to form a cluster)
τ maximum distance between two instances to be considered neighbors
Output:
Required clusters
Initialization:
Initialize six empty clusters.
$\forall$ cluster, maintain a list of ι
Create an empty set $\mathbb{E}$ to keep track of processed ι.

Calculate the Euclidean distance between two instances (ι_1, ι_2).
$\forall\, \iota \in \mathbb{D}$ Calculate reliable co-efficient $r_i(\iota_{maj}, \iota_{min})$ by using equation no 9.
$\forall \iota \in \mathbb{D}$ Calculate core distance by using equation 3.
$\forall$ unprocessed $\iota \in \mathbb{D}$:
  Calculate $\gamma(\iota_1, \iota_2)$ using equation 11.
  Add ι to $\mathbb{E}$.
$\forall \iota \in \mathbb{D}$
if $\iota \in \iota_{maj}$ **then**
  if $\gamma(\iota) < \tau$ **then**
    if $r_i(\iota) < \tau$ **then**
  Add ι to the "Overlapped majority instances" cluster.
    else
  Add ι to the "Majority instances" cluster.
      if $\gamma(\iota) > \tau$ **then**
Add ι to the "Majority noise" cluster.
      end if
    end if
  end if
end if
if $\iota \in \iota_{min}$ **then**
  if $\gamma(\iota) < \tau$ **then**
    if $r_i(\iota) < \tau$ **then**
  Add ι to the "Overlapped minority instances" cluster.
    else
  Add ι to the "Minority instances" cluster.
      if $\gamma(\iota) > \tau$ **then**
  Add ι to the "Minority noise" cluster.
      end if
    end if
  end if
end if
Return required clusters

Figure 5 illustrates the flowchart of the proposed data clustering approach and figure 6 illustrates the clusters obtained by the modified OPTICS algorithm from an overlapped and noisy imbalanced data set. In addition to being clustered together, these clusters are also distinguished by how instances of the majority and minority classes are distributed within them. To allow the proposed approach to discriminate between majority and minority noise, overlapping majority and minority instances, and non-overlapped instances of both classes, reachability distances are evaluated in a fashion that reflects the underlying distribution of classes.

Algorithm 2 Proposed Ok-NB Method for Classification

Input:
New instance $\iota_n \in$ Test data-set $\mathbb{D}_{\approx}$
Clusters/Classes created by algorithm 1
User-defined value for k
Output:
Assigned class for instance ι_n
Initialization:
Initialize a buffer $\mathbb{P}$ to store probabilities $P(\zeta)$ for each class.
Initialize a buffer $\mathbb{W}$ to store the weighted score $f(\iota)$ for each class.

$\forall\, \iota_n$ Calculate prior probability $P(\zeta)$ for ι_n with respect to each class using equation 7 by considering relevant features.
  Store $P(\zeta)$ in $\mathbb{P}$ for the class.
Calculate the class-conditional probability $P(\iota_n|\zeta)$ for the features of instance ι_n in class ζ using the equation 8.
$\forall\, \iota_n$ Calculate the weighted score $W(\iota_n)$ for each class based on the $P(\zeta)$ using equation 10.
  Store $W(\iota_n)$ in $\mathbb{W}$.
  Sort $f(\iota_n)$ in $\mathbb{W}$ in descending order.
  Select the top k weighted scores with maximum $W(\iota_n)$.

if There is a clear majority among the top *k* clusters **then**
  Assign ι_n to the cluster with the majority weighted score. And Display the assigned cluster as the classification result.
else
  if There is no clear majority or a tie among the top k clusters **then**
  Display a "No Consensus" or "Ambiguous" result to indicate the uncertainty.
  end if
end if

B. CLASSIFICATION USING PROPOSED K-NAIVE BAYES APPROACH

After the data clustering stage, the procedure moves on to the final classification task. A probabilistic approach called Naive Bayes [33] is employed for the classification process. It is renowned for its ease of use and effectiveness in processing continuous or categorical data. Features are selected based on their significance and ability to support precise classification. The preprocessed clustered data created in the previous step are used to train this proposed Ok-NB approach. Initially, it determines the prior probability $P(\zeta)$ for each cluster, which is the likelihood that a given data point will

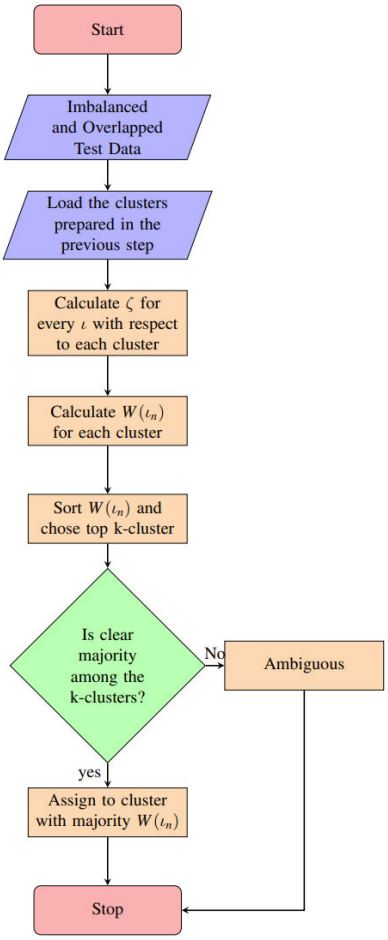

FIGURE 7. Flowchart of the proposed Ok-NB approach.

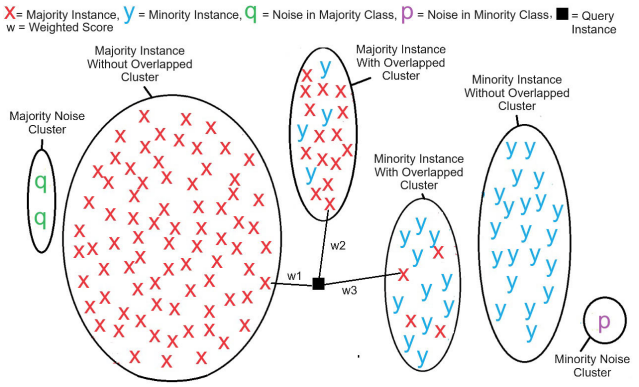

FIGURE 8. Ok-NB method.

belong to that cluster, given the total number of instances in that cluster. For each feature of the new instance ι_n in each cluster, the class-conditional probability $P(\iota_n|\zeta)$ is calculated using the Naive Bayes algorithm. Each cluster's weighted score, $W(\iota_n)$, is determined by using equation 10. Importantly, in the situation of imbalanced and overlapping datasets, features are chosen according to their capacity to facilitate accurate classification. This guarantees that the classifier is strong and able to manage the complexity found in these kinds of datasets.

After that, the weighted scores of each cluster are sorted, and then the top k clusters are chosen. The number of top-rank values is specified by the variable k. If there is a clear majority among the top k clusters, the instance is assigned to the cluster with the majority weighted score, which is the classification result. The approach shows doubt by reporting "No Consensus" or "Ambiguous" if there is no clear majority or tie among the top k clusters. This method addresses scenarios when there might not be a single dominant cluster for the new instance and permits probabilistic categorization based on weighted scores. With this approach, classification may be done in a customizable manner. The user can change the variable k to designate how many of the maximum scores should be taken into account when calculating the average score for each cluster, ensuring that the most informative instance has the most impact on the choice. The algorithm for Proposed Ok-NB is mentioned in the algorithm 2.

Figure 7 illustrates the flow chart of the proposed approach, and Figure 8 shows how the O-kNB method works. This diagram illustrates the six clusters that were produced in the earlier phase. The 'New Queries' instance is indicated by the black box. To determine which cluster is appropriate for the new instance, the Ok-NB first determines the likelihood for each cluster. Subsequently, the algorithm determines the weight of likelihood for every cluster. Following that, it will take into account the top k weight and use the majority weight selection to determine the outcome.

V. EXPERIMENTS

In this section, an attempt has been made to outline the experimental setup and process for comparing the performance of the proposed Ok-NB approach, with several well-known state-of-the-art methods. The performance of the suggested approach is assessed in three different categories: In one category the comparison is made with a few recently proposed algorithm-level approaches which are discussed above in section-II like, K-RNN [44], SVM-OTHR [45], BQI-GSA [46], GFRNN [47], NPC [48], LS-KWMTSVM [49], and DBANN [50]. It will determine whether the suggested strategy can compete with current developments in the field while maintaining its applicability and effectiveness in modern settings.

To evaluate the performance of the suggested approach with different resampling techniques, in the next category, the proposed approach is compared with a few resampling strategies like SMOTE [36], B-SMOTE [37], K-SMOTE [38], SVM-SMOTE [39], and ADASYN [40]. Here Support Vector Machine (SVM) is used as the classifier.

To establish a benchmark, finally, the comparison is done with the traditional classifier like k-nearest neighbours

TABLE 2. Details of used datasets.

Dataset	IR	Total Instance	Features	No of Minority Instances	No of Majority Instances
Ecoli1	3.36	336	8	77	259
Cleverland	3.45	303	12	68	235
Yeast2vs4	9.08	514	8	51	463
Glass015vs2	9.12	172	9	17	155
Yeast0256vs3789	9.14	1105	8	109	996
Ecoli0267vs35	9.18	224	7	22	202
Yeast05679vs4	9.35	528	8	51	477
Vowel0	9.98	989	13	90	899
Glass016vs2	10.29	192	9	17	175
Glass2	11.59	214	9	17	197
Cleveland0vs4	12.62	177	13	13	164
Shuttlec0vsc4	13.87	1889	9	127	1762
Glass4	15.46	214	9	13	201
Page blocks13vs4	15.86	472	10	28	444
Abalone918	16.4	731	8	42	689
Shuttlec2vsc4	20.5	129	9	6	123
Glass5	22.78	214	9	9	205
Yeast4	28.1	1484	8	51	1433
Yeast5	32.73	1484	8	44	1440
Yeast6	41.4	1484	8	35	1449
Abalone19	129.44	4174	8	32	4142

(k-NN) [41], Support Vector Machine(SVM) [42] and Random Forest (RF) [43].

The main objective of this comparison is to evaluate the advantages, disadvantages, and overall efficacy in real-world applications of this proposed approach. This is done by thoroughly comparing the proposed approach with recent algorithm-level approaches, resampling techniques, and conventional classifiers using a variety of datasets with varying degrees of class imbalance and real-world settings. To ensure uniformity and fairness in comparisons, every experiment is carried out in a setting with the same hardware and software. The KEEL [61] open-source platform is used for the development of all comparison methodology programs as well as learning tools.

A. BENCHMARK DATASET

A detailed experiment has been done on 21 binary-class imbalanced datasets taken from the KEEL [62] data set repository to evaluate the effectiveness of this suggested strategy. These datasets were frequently used to assess the effectiveness of various techniques. The range of the imbalance ratios is 3.36 to 129.44. Table 2 lists the specifics of the descriptions of the experimental dataset.

B. PARAMETER SETTING

For performance evaluation, the proposed method, Ok-NB is compared with other approaches that include state of art classifiers like k-NN [41], SVM [42], RF [43] and a few advanced algorithm level approach like K-RNN [44], SVM-OTHR [45], BQI-GSA [46], GFRNN [47], NPC [48], LS-KWMTSVM [49], and DBANN [50]. The parameter k is picked from the original literature and set to 3, 3, and 1, respectively, for the kNN-based methods kNN, DBANN, and kRNN. According to the original literature, the other parameter like Minpts is set to 4 for DBANN, and 10-fold cross-validation is used to select eps as the best value from the range [0.01, 200]. For SVM and Ada-SVM C, degree and gamma are considered 1.0, 3, and auto, respectively.

The OPTICS base clustering requires two input parameters, which are the minimum number of points to form a cluster (MinPts) and the maximum distance between two instances to consider neighbors (τ), respectively. According to earlier studies, (MinPts) has minimal effect on the clustering outcomes [32]. For this reason, to examine its effects on clustering performance, (MinPts) is set to 5, and (τ) is adjusted between 0.01 and 0.1 in this phase. Three real-world datasets were employed in these experiments, and the outcomes are listed in a table 3. Figure 9 demonstrates that (τ) is a sensitive parameter that greatly influences the clustering performance.

The number of clusters tends to decrease as the (τ) value increases, and the noise level also tends to decrease. This shows that less clustering occurs when (τ) values are bigger. Furthermore, it seems that the influence of (τ) on the number of clusters and noise levels differs among datasets, suggesting possible differences in data properties and clustering efficacy. Although the three datasets had varying noise levels, the suggested approach eventually produced the required number of clusters at (τ) = 0.05. According to this, (τ) = 0.05 would be a good value to use to get the desired clustering outcome for subsequent tasks.

Finally, for the proposed Ok-NB, the threshold values for (MinPts) and (τ) are considered 5 and 0.05, respectively. The parameter k controls the number of clusters assessed for categorization. It can be any value between 1 and 6. The accuracy of the classification results is directly impacted by the parameter k selection. A lower value of k could lead to a more limited representation of the cluster and possibly a more skewed result. On the other hand, a higher value of

TABLE 3. Clustering results based on different (τ) values.

Dataset	Yeast2vs4		Yeast0256vs3789		Glass016vs2	
(τ)	No of Cluster	Noise	No of Cluster	Noise	No of Cluster	Noise
0.01	12	60	18	111	9	31
0.02	10	50	15	80	8	25
0.03	8	32	13	75	6	21
0.04	8	28	9	53	6	17
0.05	6	20	6	41	6	11
0.06	5	11	4	25	4	9
0.07	3	5	3	15	3	5
0.08	3	1	3	7	3	0
0.09	2	0	2	0	2	0
0.1	3	0	2	0	2	0

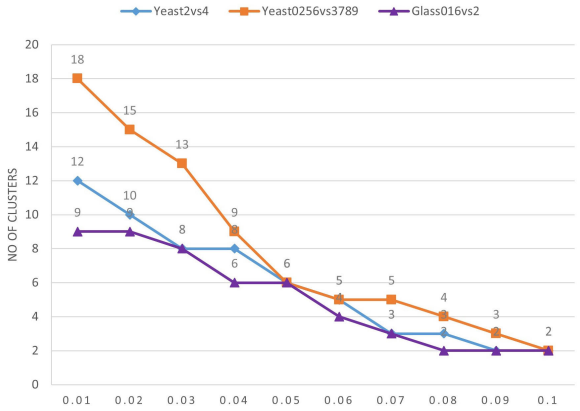

FIGURE 9. Clustering analysis with different(τ) values.

k might contain more clusters, which would allow the data to show more nuances results. The possibility of running across ambiguous instances is influenced by sensitivity to parameter k. Due to restricted cluster representation, a smaller value of k might lead to more instances being categorized as ambiguous; conversely, a larger value of k might decrease the frequency of ambiguous classifications but increase computational cost. Computational efficiency is also affected by the selection of parameter k. A higher value of k requires more computational time and resources. Conversely, a lower value of k minimizes processing cost, but it may result in worse classification accuracy by ignoring important clusters. For the final classification, the k is considered 3.

1) PERFORMANCE METRICS

AUC (Area Under the Receiver Operating Characteristic Curve), F1 Score, and Accuracy are the evaluation measures used to assess the efficacy of this suggested approach, Ok-NB. AUC is especially important when contrasting the ability of a suggested strategy to discriminate against alternative approaches. It analyzes how effectively a model can discriminate between positive and negative instances [63]. In situations where there is a class imbalance, the F1-score becomes very important. It strikes a balance between recall and precision. It considers the significance of false positives as well as false negatives in analysis [63]. A broad indicator of overall correctness in some classification outcomes is accuracy. It is essential to understand the percentage of accurately predicted events in every class. It is possible to evaluate a given approach's efficiency in producing accurate predictions throughout the whole dataset by comparing its accuracy with the accuracy of other approaches [63]. The main objective is to offer a thorough assessment of the model's classification performance for the specified task. These three measurements can offer a complete picture of how well the suggested technique performs in the classification challenge. They enable us to assess the model's discriminative power (AUC) [63], precision-to-recall trade-off (F1 Score) [63], and overall prediction accuracy (Accuracy) [63]. By using these measures, we want to offer a comprehensive evaluation of the model's effectiveness and applicability for the desired use.

$$\text{True} - \text{Positive} - \text{Rate(TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

$$\text{False} - \text{Positive} - \text{Rate(FPR)} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{13}$$

$$\text{AUC} = \frac{1 + \text{TPR} - \text{FPR}}{2} \tag{14}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{15}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{16}$$

$$\text{F1} - \text{Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \cdot 100 \tag{18}$$

where TP, FP, TN, and FN are true positive, false positive, true negative, and false negative, respectively.

VI. RESULT AND DISCUSSION

A few comparable experiments on all of the datasets listed in Table 2 with the other imbalanced classification approaches described above have been done in order to confirm the effectiveness of the proposed approach in handling imbalanced datasets with different imbalance ratios. For each experiment, the AUC, F1-score, and accuracy have been calculated. With respect to each and every data set, the AUC, F1-score, and accuracy rate of all the recently proposed

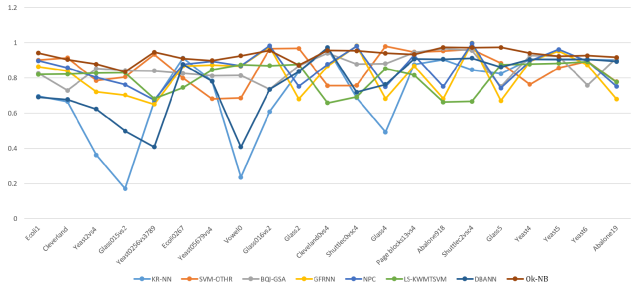

FIGURE 10. The average AUC of various approaches over the dataset.

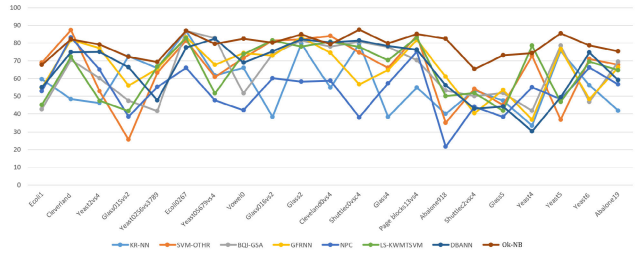

FIGURE 11. The average F1-Score (in%) of various approaches over the dataset.

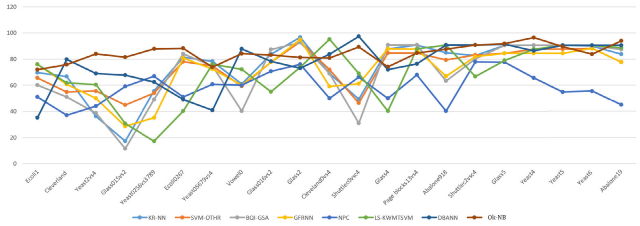

FIGURE 12. The average accuracy (in%) of various approaches over the dataset.

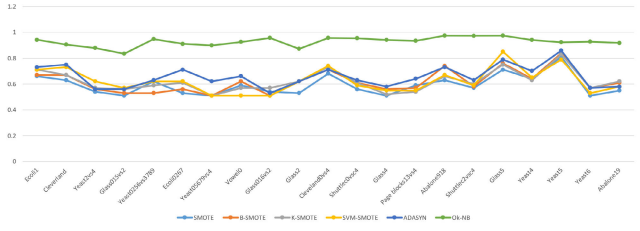

FIGURE 13. The average AUC of various resampling approaches over the dataset.

algorithm-level approaches are listed in the table 4, 5, and 6 respectively. Similarly, the AUC, F1-score, and accuracy rate of various resampling approaches using the SVM as a base classifier are listed in table 7, 8, and 9, respectively. The table 10 consists of the same values for the traditional classifier.

Figures 10, 11, and 12 present the graphical representation of the performance metrics - AUC, F1 score, and Accuracy rate, respectively, for the recent algorithm-level approaches. These figures illustrate how well the proposed approach performs across different evaluation criteria over

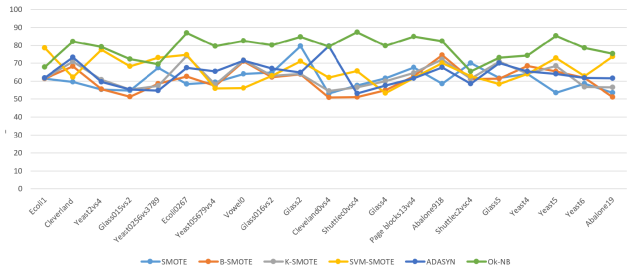

FIGURE 14. The average F1-score (in%) of various resampling approaches over the dataset.

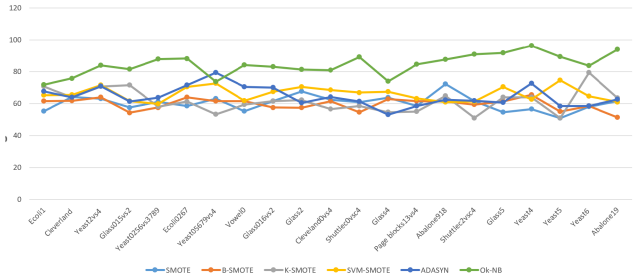

FIGURE 15. The average accuracy (in%) of various resampling approaches over the dataset.

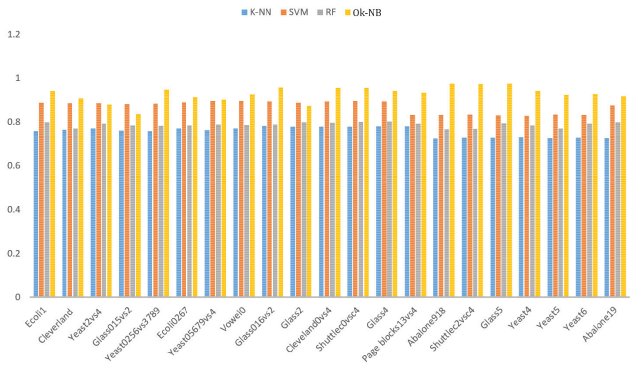

FIGURE 16. The average AUC of various traditional classifiers and OK-NB over the dataset.

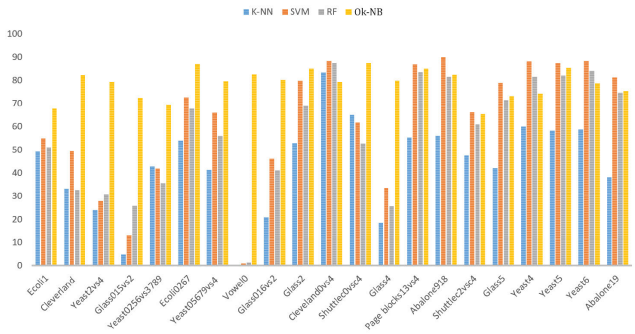

FIGURE 17. The average F1-Score (in%) of various traditional classifiers and OK-NB over the dataset.

the recent algorithm-level approaches. The AUC, F1 score, and accuracy rate for the various resampling approaches are shown graphically in Figures 13, 14, and 15 respectively.

TABLE 4. The average AUC of various approaches over the dataset.

Dataset	KR-NN	SVM-OTHR	BQI-GSA	GFRNN	NPC	LS-KWMT-SVM	DBANN	Ok-NB
Ecoli1	0.696	0.902	0.826	0.864	0.898	0.821	0.691	0.942
Cleverland	0.666	0.915	0.729	0.841	0.858	0.823	0.678	0.905
Yeast2vs4	0.362	0.786	0.854	0.722	0.803	0.828	0.623	0.879
Glass015vs2	0.171	0.806	0.843	0.703	0.762	0.83	0.498	0.835
Yeast0256vs3789	0.665	0.934	0.841	0.651	0.676	0.683	0.408	0.947
Ecoli0267	0.914	0.799	0.826	0.868	0.878	0.746	0.877	0.911
Yeast05679vs4	0.783	0.682	0.813	0.874	0.894	0.847	0.782	0.899
Vowel0	0.236	0.687	0.815	0.867	0.869	0.875	0.408	0.926
Glass016vs2	0.609	0.967	0.736	0.979	0.984	0.87	0.735	0.957
Glass2	0.837	0.968	0.87	0.681	0.753	0.877	0.837	0.873
Cleveland0vs4	0.966	0.756	0.941	0.868	0.878	0.658	0.973	0.956
Shuttlec0vsc4	0.689	0.757	0.879	0.983	0.981	0.692	0.719	0.955
Glass4	0.492	0.98	0.882	0.683	0.751	0.854	0.763	0.941
Page blocks13vs4	0.877	0.948	0.946	0.869	0.93	0.817	0.908	0.934
Abalone918	0.905	0.954	0.967	0.683	0.752	0.663	0.907	0.974
Shuttlec2vsc4	0.847	0.962	0.956	0.998	0.996	0.667	0.913	0.973
Glass5	0.824	0.8843	0.752	0.671	0.743	0.875	0.862	0.974
Yeast4	0.906	0.763	0.927	0.895	0.897	0.879	0.906	0.941
Yeast5	0.904	0.857	0.923	0.953	0.962	0.884	0.906	0.924
Yeast6	0.904	0.894	0.758	0.874	0.894	0.892	0.906	0.927
Abalone19	0.902	0.779	0.914	0.681	0.753	0.777	0.894	0.918

TABLE 5. The average F1-Score (in%) of various approaches over the dataset.

Data-set	KR-NN	SVM-OTHR	BQI-GSA	GFRNN	NPC	LS-KWMT-SVM	DBANN	Ok-NB
Ecoli1	59.67	68.9	42.63	54.74	52.83	45.06	55.02	67.8
Cleverland	48.33	87.17	70.57	81.67	83.16	72.16	74.81	82.12
Yeast2vs4	46.01	52.71	60.11	77.02	65.00	47.33	75.03	79.11
Glass015vs2	72.59	25.66	47.33	56.00	38.4	41.53	66.33	72.31
Yeast0256vs3789	65.93	63.33	41.53	65.53	55.14	66.67	47.66	69.34
Ecoli0267	86.67	81.33	86.67	83.33	66.02	82.67	77.33	86.88
Yeast05679vs4	61.67	60.89	82.67	67.8	47.52	51.57	82.33	79.54
Vowel0	66.10	71.48	51.57	74.33	42.05	73.71	69.00	82.44
Glass016vs2	38.33	81.33	73.71	72.88	60.10	81.33	75.36	80.12
Glass2	79.67	82.00	81.33	83.33	58.24	78.00	82.05	84.78
Cleveland0vs4	54.8	84.00	78.00	74.46	58.76	80.67	80.33	79.23
Shuttlec0vsc4	79.67	74.61	80.67	56.76	38.02	77.81	81.33	87.34
Glass4	38.33	66.05	77.81	64.74	57.26	70.33	78.22	79.76
Page blocks13vs4	54.80	84.13	70.33	81.67	75.19	83.07	76.23	84.95
Abalone918	40.00	35.00	53.07	60.97	21.67	50.00	56.26	82.34
Shuttlec2vsc4	53.33	54.09	50.00	40.38	43.94	51.69	42.91	65.43
Glass5	47.39	45.00	51.69	53.43	38.33	41.67	44.19	73.12
Yeast4	33.42	72.67	41.67	36.67	55.03	78.5	30.44	74.23
Yeast5	75.53	36.67	78.5	76.47	48.33	46.67	49.33	85.27
Yeast6	56.15	70.97	46.67	48.33	66.28	69.42	74.6	78.55
Abalone19	41.81	67.8	69.42	66.28	56.76	64.74	59.2	75.32

Compared to the resampling strategy, these graphs help to understand how well the suggested approach performs across several evaluation criteria. The AUC, F1 score, and accuracy rate for the traditional classifiers are shown graphically in Figures 16, 17, and 18, respectively. The performance of the suggested approach in comparison to the traditional classifiers is demonstrated by these figures across a variety of evaluation criteria.

Based on the AUC, F1-Score, and Accuracy, the performance of the proposed Ok-NB approach across several datasets in the context of addressing imbalanced data is shown in Table 4, 5, 6, 7, 8, 9 and 10. By addressing several aspects of classification assessment, the combination of AUC, Accuracy, and F1-score offers a comprehensive assessment of the performance of the suggested approach. Good discriminating ability is shown by a high AUC value, overall correctness is indicated by high accuracy, and the evaluation is robust across several dimensions when the F1-score optimizes precision and recall.

The AUC value of Ok-NB varies based on the data set. It receives competitive AUC values on large datasets like Ecoli1 and Yeast5, demonstrating its efficacy in these situations. On other datasets like Cleveland0vs4 and Vowel0, the performance is average. It is able to efficiently handle

TABLE 6. The average accuracy (in%) of various approaches over the dataset.

Data-set	KR-NN	SVM-OTHR	BQI-GSA	GFRNN	NPC	LS-KWMT-SVM	DBANN	Ok-NB
Ecoli1	69.61	65.55	60.16	75.89	51.03	75.98	35.15	71.88
Cleverland	66.60	54.82	50.95	60.68	37.01	61.75	79.72	75.82
Yeast2vs4	36.27	55.56	38.68	49.96	44.02	60.43	69.01	84.01
Glass015vs2	17.13	44.79	11.42	28.55	58.96	30.64	67.80	81.51
Yeast0256vs3789	55.56	54.02	49.08	35.15	66.89	17.13	62.39	87.85
Ecoli0267	81.42	78.1	83.98	80.89	51.03	40.27	49.08	88.22
Yeast05679vs4	78.36	74.79	74.70	72.40	60.68	76.18	40.83	73.79
Vowel0	60.98	59.21	40.21	60.08	60.00	72.25	87.75	84.21
Glass016vs2	83.79	77.58	87.47	77.58	70.65	55.12	78.22	83.11
Glass2	96.65	93.29	92.59	95.01	75.89	73.79	73.05	81.35
Cleveland0vs4	68.96	71.96	68.97	58.96	50.00	95.33	83.78	80.88
Shuttlec0vsc4	49.26	46.36	30.87	61.23	66.15	68.97	97.39	89.28
Glass4	87.76	84.69	90.78	87.76	50.00	40.27	71.97	74.02
Page blocks13vs4	90.75	84.58	90.70	87.66	67.85	87.76	76.39	84.55
Abalone918	84.76	79.28	63.35	66.89	40.27	90.75	90.80	87.62
Shuttlec2vsc4	82.84	83.09	80.25	81.84	77.98	66.58	90.74	90.86
Glass5	90.66	84.42	90.59	84.42	77.64	78.86	91.37	91.75
Yeast4	90.64	87.53	90.59	84.4	65.55	87.54	86.28	96.44
Yeast5	90.64	87.51	90.59	84.39	54.82	90.64	90.65	89.48
Yeast6	90.02	88.23	90.37	88.14	55.56	90.64	90.63	83.78
Abalone19	83.79	77.58	87.47	77.58	44.79	88.71	90.62	93.91

TABLE 7. The average AUC of various resampling approaches over the dataset.

Data-set	SMOTE	B-SMOTE	K-SMOTE	SVM-SMOTE	ADASYN	Ok-NB
Ecoli1	0.66	0.67	0.71	0.71	0.73	0.942
Cleverland	0.63	0.67	0.67	0.73	0.75	0.905
Yeast2vs4	0.54	0.56	0.57	0.62	0.56	0.879
Glass015vs2	0.51	0.53	0.56	0.57	0.56	0.835
Yeast0256vs3789	0.62	0.53	0.59	0.62	0.63	0.947
Ecoli0267	0.53	0.56	0.61	0.62	0.71	0.911
Yeast05679vs4	0.51	0.51	0.51	0.51	0.62	0.899
Vowel0	0.59	0.62	0.57	0.51	0.66	0.926
Glass016vs2	0.54	0.51	0.57	0.51	0.53	0.957
Glass2	0.53	0.62	0.62	0.62	0.62	0.873
Cleveland0vs4	0.68	0.71	0.73	0.74	0.71	0.956
Shuttlec0vsc4	0.56	0.61	0.61	0.59	0.63	0.955
Glass4	0.51	0.56	0.52	0.55	0.58	0.941
Page blocks13vs4	0.59	0.57	0.54	0.55	0.64	0.934
Abalone918	0.63	0.74	0.66	0.67	0.73	0.974
Shuttlec2vsc4	0.57	0.58	0.6	0.59	0.63	0.973
Glass5	0.71	0.76	0.75	0.85	0.79	0.974
Yeast4	0.64	0.64	0.63	0.65	0.7	0.941
Yeast5	0.81	0.84	0.83	0.79	0.86	0.924
Yeast6	0.51	0.57	0.57	0.53	0.57	0.927
Abalone19	0.55	0.61	0.62	0.58	0.58	0.918

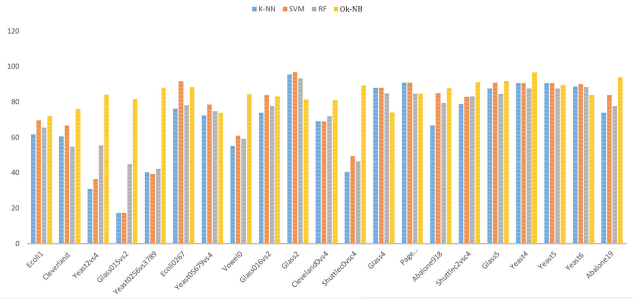

FIGURE 18. The average accuracy (in%) of various traditional classifiers and OK-NB over the dataset.

skewed data under any circumstances. On datasets like Yeast4, Ecoli0267, and Glass016vs2, Ok-NB maintains comparatively constant performance with AUC values in the mid to high range. This consistency can be a sign of its dependability for particular classes of imbalanced datasets. On datasets with overlapping data, like Glass5 and Yeast6, when the AUC values are good, this indicates the technique performs better. This shows that Ok-NB is capable of effectively separating classes that overlap. On various datasets, Ok-NB often achieves a balance between precision and recall. It keeps competitive performance while avoiding excessive overfitting to the majority class, which can be a problem in circumstances with imbalanced data.

The variability across several datasets is seen by the F1-Score values for this proposed Ok-NB method. In some datasets, like Cleveland0vs4 and Yeast5, the approach

TABLE 8. The average F1-Score of various resampling approaches over the dataset.

Data-set	SMOTE	B-SMOTE	K-SMOTE	SVM-SMOTE	ADASYN	Ok-NB
Ecoli1	61.41	61.42	61.52	78.42	61.73	67.8
Cleverland	59.46	68.25	70.75	62.25	73.22	82.12
Yeast2vs4	55.23	55.47	60.77	77.58	59.46	79.11
Glass015vs2	54.66	51.31	55.35	68.21	55.23	72.31
Yeast0256vs3789	67.48	58.62	57.34	73.08	54.66	69.34
Ecoli0267	58.42	62.65	74.03	74.75	67.48	86.88
Yeast05679vs4	59.41	57.42	58.07	55.86	65.48	79.54
Vowel0	64.05	70.99	71.58	56.02	71.52	82.44
Glass016vs2	64.8	62.16	62.87	62.98	66.91	80.12
Glass2	79.5	63.72	63.99	71.12	64.8	84.78
Cleveland0vs4	53.08	50.94	54.6	61.92	79.5	79.23
Shuttlec0vsc4	57.54	50.99	56.51	65.54	53.08	87.34
Glass4	61.52	54.97	59.88	53.29	57.54	79.76
Page blocks13vs4	67.61	62.73	64.61	61.85	61.52	84.95
Abalone918	58.6	74.56	72.14	70	67.61	82.34
Shuttlec2vsc4	70.06	61.21	61.3	62.81	58.6	65.43
Glass5	61.66	61.39	70.78	58.41	70.06	73.12
Yeast4	63.99	68.47	64.26	63.89	65.42	74.23
Yeast5	53.4	65.42	68.47	72.89	63.99	85.27
Yeast6	58.57	61.79	56.67	62.81	61.72	78.55
Abalone19	53.53	51.03	56.59	73.74	61.56	75.32

TABLE 9. The average accuracy of various resampling approaches over the dataset.

Data-set	SMOTE	B-SMOTE	K-SMOTE	SVM-SMOTE	ADASYN	Ok-NB
Ecoli1	55.23	61.52	70.78	65.23	67.61	71.88
Cleverland	64.26	61.72	63.99	65.42	63.99	75.82
Yeast2vs4	63.08	64.05	70.75	71.58	70.99	84.01
Glass015vs2	57.54	54.28	71.52	61.66	61.39	81.51
Yeast0256vs3789	61.21	57.54	58.07	59.88	63.84	87.85
Ecoli0267	58.57	63.99	61.25	70.42	71.52	88.22
Yeast05679vs4	63.17	61.66	53.4	72.75	79.5	73.79
Vowel0	55.35	61.56	59.46	61.79	70.46	84.21
Glass016vs2	61.52	57.54	61.66	67.48	70.06	83.11
Glass2	67.61	57.42	62.16	70.46	60.42	81.35
Cleveland0vs4	62.33	61.52	56.51	68.47	64.26	80.88
Shuttlec0vsc4	61.01	54.66	58.6	66.91	61.31	89.28
Glass4	63.72	62.87	54.66	67.48	53.08	74.02
Page blocks13vs4	58.62	61.52	54.97	63.22	58.6	84.55
Abalone918	72.14	61.42	64.8	61.01	62.33	87.62
Shuttlec2vsc4	61.52	59.41	51.03	61.25	61.79	90.86
Glass5	54.6	61.39	63.99	70.42	60.77	91.75
Yeast4	56.59	65.48	63.99	62.73	72.75	96.44
Yeast5	50.99	54.97	50.94	74.56	58.42	89.48
Yeast6	58.07	58.62	79.5	64.66	58.57	83.78
Abalone19	61.52	51.31	63.61	61.03	62.65	93.91

performs exceptionally well and receives high F1-Scores. In contrast, it performs an average in other datasets like Yeast4 and Glass015vs2, where its F1-Scores are relatively lower. In datasets where it performs well, it demonstrates how well it can classify. The performance of this method seems to be consistent. As evidenced by the high F1-Scores for ''Glass015vs2'' and ''Yeast2vs4'', the approach performs better in datasets with overlapped data. This shows that Ok-NB can efficiently discriminate between classes that are overlapped.

Across many datasets, the OK-NB method's accuracy shows a large amount of variability. In some instances, it achieves excellent accuracy, like in ''Glass2'' and ''Shuttlec0vsc4''. However, in datasets like ''Glass015vs2'' and ''Yeast2vs4'', where accuracy is average, its performance is average. High accuracy numbers show that it can accurately categorize instances in datasets where it excels. This demonstrates its capacity to handle skewed data well in particular contexts. The sensitivity to dataset characteristics is also better than the other approach. This implies that it performs better than alternative methods.

The experimental findings indicate that Ok-NB outperforms recently suggested algorithm-level approaches. The AUC, F1 score, and accuracy metrics demonstrate its improved performance, which highlights how well it can address class imbalance and overlap issues to improve classification accuracy. Therefore, the proposed approach emerges as a promising solution for handling imbalanced

TABLE 10. The average AUC, F1-Score (in%) and accuracy (in%) of various approaches over the dataset.

Dataset	AUC				F1-Score				Accuracy			
	K-NN	SVM	RF	Ok-NB	K-NN	SVM	RF	Ok-NB	K-NN	SVM	RF	Ok-NB
Ecoli1	0.758	0.886	0.798	0.942	49.21	54.80	50.93	67.80	61.75	69.61	65.55	71.88
Cleverland	0.763	0.884	0.770	0.905	33.09	49.30	32.51	82.12	60.43	66.60	54.82	75.82
Yeast2vs4	0.770	0.884	0.791	0.879	23.88	27.71	30.76	79.11	30.64	36.27	55.56	84.01
Glass015vs2	0.76	0.88	0.784	0.836	4.86	13.00	25.78	72.31	17.13	17.13	44.79	81.51
Yeast0256vs3789	0.758	0.882	0.781	0.947	42.64	41.81	35.53	69.34	40.00	39.22	41.98	87.85
Ecoli0267	0.769	0.889	0.784	0.911	53.91	72.59	67.73	86.88	76.18	91.42	78.10	88.22
Yeast05679vs4	0.762	0.894	0.787	0.899	41.30	65.93	56.04	79.54	72.25	78.36	74.79	73.79
Vowel0	0.770	0.894	0.786	0.926	0.00	0.98	1.33	82.44	55.12	60.98	59.21	84.21
Glass016vs2	0.781	0.892	0.787	0.957	20.83	46.00	41.05	80.12	73.79	83.79	77.58	83.11
Glass2	0.778	0.886	0.798	0.873	52.83	79.67	68.90	84.78	95.33	96.65	93.29	81.35
Cleveland0vs4	0.778	0.891	0.795	0.956	83.16	88.31	87.17	79.23	68.97	68.96	71.96	80.88
Shuttlec0vsc4	0.778	0.894	0.799	0.955	65.00	61.67	52.71	87.34	40.27	49.26	46.36	89.28
Glass4	0.780	0.891	0.801	0.941	18.40	33.42	25.66	79.76	87.76	87.76	84.69	74.02
Page blocks13vs4	0.779	0.831	0.791	0.934	55.14	86.67	83.33	84.95	90.75	90.75	84.58	84.55
Abalone918	0.723	0.830	0.766	0.974	56.02	90.00	81.33	82.34	66.58	84.76	79.28	87.62
Shuttlec2vsc4	0.726	0.832	0.768	0.973	47.52	66.10	60.89	65.43	78.86	82.84	83.09	90.86
Glass5	0.727	0.828	0.793	0.974	42.05	78.87	71.48	73.12	87.54	90.66	84.42	91.75
Yeast4	0.728	0.826	0.784	0.941	60.10	88.00	81.33	74.23	90.64	90.64	87.53	96.44
Yeast5	0.725	0.832	0.769	0.924	58.24	87.33	82.00	88.27	90.64	90.64	87.51	89.48
Yeast6	0.727	0.830	0.791	0.927	58.76	88.33	84.00	78.55	88.71	90.02	88.23	83.78
Abalone19	0.725	0.875	0.798	0.918	38.02	81.11	74.61	75.32	73.79	83.79	77.58	93.91

as well as overlapped data across various domains, offering a robust and reliable approach. Regarding accuracy, F1 score, and AUC, it outperforms other resampling approaches. Because of its intrinsic algorithmic architecture, it can obtain improved classification results without requiring the creation of artificial data or the alteration of already-existing samples. Compared to conventional classifiers, it also performs better on all metrics. The class imbalance problems can be effectively tackled by its unique combination of probabilistic classification and the modified OPTICS algorithm.

A. PERFORMANCE ON REAL-WORLD DATASET

In a variety of real-world application contexts, the Ok-NB method shows potential, especially in fields where classification tasks face difficulties due to imbalanced data. Its efficiency in tackling certain problems in these domains is highlighted by its performance, which can be observed in the tables 4, 5, 6, 7, 8, 9 and 10 that illustrate its outcomes across various datasets. In datasets such as Ecoli1 and Yeast5, the Ok-NB algorithm's competitive AUC values show how well it can classify occurrences, which helps with disease diagnosis and prognosis. It can handle imbalanced biomedical datasets with reliability, as evidenced by its consistent performance on a variety of datasets, including Yeast4 and Glass016vs2. It can also distinguish between classes that overlap, as demonstrated by its strong results on datasets such as Yeast6 and Glass5, which makes it appropriate for precisely detecting fraudulent transactions. Furthermore, its competitive accuracy highlights its potential in fraud detection applications, especially in datasets like Shuttlec0vsc4 and Shuttlec2vsc4. The remarkably constant performance of the algorithm on datasets such as Ecoli0267 and Glass4 indicates that it is a suitable method for accurately identifying defective products. Its relevance in industrial quality control tasks is further enhanced by its robustness in striking a compromise between precision and recall, as shown across many datasets.

B. TIME COMPLEXITY ANALYSIS

Although the main goal of the paper is to achieve efficient imbalance classification, we present a method where the time complexity of OPTICS is enhanced by providing the required clusters obtained with modified readability distance with overlapping instances only. Thus proposed method reduces the run time to some extent.

The number of classes (C) and the amount of k are the two key factors that determine the time complexity of the proposed Ok-NB strategy for classification. It performs well, with an O(ClogC) average and worst-case time complexity, where C is the number of classes. For real-time classification tasks, the technique is appropriate since its time complexity is typically minimal and independent of the amount of the training dataset. The Ok-NB method offers a more straightforward and computationally efficient way of classification compared to other algorithms like k-nearest neighbours (KNN) or support vector machines (SVM). This is especially true when working with high-dimensional data. It works well for real-time classification jobs where immediate choices are necessary because of its low temporal complexity and high computational efficiency.

VII. CONCLUSION AND FUTURE WORKS

It is particularly challenging to appropriately categorize an instance because of the presence of imbalance and overlap data in the training dataset. This study offered a novel approach by combining the concepts of clustering and classification. It combines the advantages of both the OPTICS and Naive-Base approach. By modifying the reachability distance

function of OPTICS, along with incorporating a new weight function and considering the top k weight in Naive-Bayes, this approach produces classification results that are more reliable and accurate in an imbalanced environment. Utilizing the power of a modified OPTICS algorithm to cluster data into identical groups and then makes use of the adaptable naive Bayes classifier for class determination based on maximum score count. The objective was to address the difficulties of overlapping and imbalanced data simultaneously and present a workable and effective solution. The outcomes of in-depth experiments and analyses confirm the viability of this suggested approach. The data are effectively divided into identical clusters that more correctly reflect the underlying structure, including regions of class overlap, by using a modified OPTICS method that is optimized for density-based clustering. The incorporation of the Naive Bayes classifier made it possible to generate a proper prediction based on a maximum score count inside these clusters, which improves the overall predictive accuracy.

This approach has several significant benefits. It simplifies the classification procedure for overlapping, imbalanced data by eliminating the requirement for complex feature engineering and resampling approaches. Additionally, the clustering feature not only helps to balance out class imbalance but also offers important insights into the fundamental structure of the data. It is crucial to understand that not every issue involving imbalanced data can be solved by using a single solution. Further study is required to examine the constraints and adaptability of this approach across other problem domains, as the inherent properties of the dataset can affect the performance of any approach. The proposed algorithm-level technique contributes significantly to the categorization of overlapped imbalanced data. In situations when the class overlap is a substantial challenge, it may improve the performance of predictive models.

Ok-NB has some limitations even though it provides an easy-to-understand method for classification jobs. To achieve its successful implementation in real-world circumstances, careful consideration of parameter selection, data pretreatment, and evaluation is required. The performance of the proposed approach may vary depending on the user-defined parameter k that is selected. Inappropriate value selection for k could result in poor classification outcomes. The computational cost of determining probabilities and weighted scores for each class may decrease the performance of the algorithm with datasets that have high-dimensional feature space.

This study is only concerned with the binary class problem. It may be possible to expand the Ok-NB approach in the future to address the multiclass problem. Further study could improve the state-of-the-art in multiclass classification and aid in the creation of a more reliable and efficient approach by tackling the particular difficulties presented by multiclass imbalanced datasets with overlapping and noisy instances.

A number of important factors and possible changes would need to be taken into account while adapting the methodology for multiclass classification, such as:

- The cluster representation
- The determination of prior probabilities for each cluster
- The class-conditional probability calculation
- The weighted score and Classification decision

This study establishes the groundwork for further investigation of hybrid algorithms that make use of clustering methods for class separation and classification schemes for accurate prediction. Future research could concentrate on creating and improving clustering and classification algorithms designed especially for noisy, overlapping, multiclass-imbalanced data sets. This involves investigating cutting-edge methods for feature selection, cluster identification, and classification judgment in multiclass situations.

The proposed method is based on clustering datasets that are imbalanced using the OPTICS algorithm and a modified reachability distance. Future work may concentrate on enhancing clustering methodologies to more effectively manage the complexity of imbalanced data sets. This could involve the investigation of alternate clustering algorithms or the integration of ensemble clustering techniques to enhance cluster quality and separation. It is assumed that this study will inspire creative and practical responses to real-world problems of overlapped imbalanced data, and open up new directions for research and applications in a variety of fields as the area of machine learning continues to develop.

REFERENCES

[1] P. Kaur and A. Gosain, "Issues and challenges of class imbalance problem in classification," *Int. J. Inf. Technol.*, vol. 14, pp. 1–7, 2018.

[2] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, "Classification with class imbalance problem," *Int. J. Advance Soft Comput. Appl.*, vol. 5, no. 3, pp. 176–204, 2013.

[3] K. Madasamy and M. Ramaswami, "Data imbalance and classifiers: Impact and solutions from a big data perspective," *Int. J. Comput. Intell. Res.*, vol. 13, no. 9, pp. 2267–2281, 2017.

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

[5] T. Hasanin, T. M. Khoshgoftaar, J. L. Leevy, and R. A. Bauder, "Severely imbalanced big data challenges: Investigating data sampling approaches," *J. Big Data*, vol. 6, no. 1, pp. 1–25, Dec. 2019.

[6] A. Fernández, S. del Río, N. V. Chawla, and F. Herrera, "An insight into imbalanced big data classification: Outcomes and challenges," *Complex Intell. Syst.*, vol. 3, no. 2, pp. 105–120, Jun. 2017.

[7] N. Rout, D. Mishra, and M. K. Mallick, "Handling imbalanced data: A survey," in *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications*. Springer, 2018, pp. 431–443.

[8] C. Lemnaru and R. Potolea, "Imbalanced classification problems: Systematic study, issues and best practices," in *Proc. Int. Conf. Enterprise Inf. Syst.* Springer, 2011, pp. 35–50.

[9] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, Nov. 2016.

[10] Z. Ahmed and S. Das, "A comparative analysis on recent methods for addressing imbalance classification," *Social Netw. Comput. Sci.*, vol. 5, no. 1, pp. 1–18, Nov. 2023.

[11] A. S. Shirkhorshidi, S. Aghabozorgi, T. Y. Wah, and T. Herawan, "Big data clustering: A review," in *Proc. Int. Conf. Comput. Sci. Appl. (ICCSA)*. Springer, 2014, pp. 707–720.

[12] Z. Ahmed, S. M. S. Askari, and S. Das, "Comparative analysis of recent data-level methods for imbalance classification," in *Proc. 4th Int. Conf. Comput. Commun. Syst. (ICS)*, Mar. 2023, pp. 1–6.

[13] L. Abdi and S. Hashemi, "To combat multi-class imbalanced problems by means of over-sampling techniques," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 1, pp. 238–251, Jan. 2016.
[14] M. Bach, A. Werner, J. Zywiec, and W. Pluskiewicz, "The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis," *Inf. Sci.*, vol. 384, pp. 174–190, Apr. 2017.
[15] Q. Wang, Y. Tian, and D. Liu, "Adaptive FH-SVM for imbalanced classification," *IEEE Access*, vol. 7, pp. 130410–130422, 2019.
[16] A. K. I. Hassan and A. Abraham, "Modeling insurance fraud detection using imbalanced data classification," in *Advances in Nature and Biologically Inspired Computing*. Springer, 2016, pp. 117–127.
[17] H. Zhu, G. Liu, M. Zhou, Y. Xie, A. Abusorrah, and Q. Kang, "Optimizing weighted extreme learning machines for imbalanced classification and application to credit card fraud detection," *Neurocomputing*, vol. 407, pp. 50–62, Sep. 2020.
[18] V. García, A. I. Marqués, and J. S. Sánchez, "Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction," *Inf. Fusion*, vol. 47, pp. 88–101, May 2019.
[19] M.-J. Kim, D.-K. Kang, and H. B. Kim, "Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction," *Exp. Syst. Appl.*, vol. 42, no. 3, pp. 1074–1082, Feb. 2015.
[20] O. M. Alyasiri, Y.-N. Cheah, and A. K. Abasi, "Hybrid filter-wrapper text feature selection technique for text classification," in *Proc. Int. Conf. Commun. Inf. Technol. (ICICT)*, Jun. 2021, pp. 80–86.
[21] O. M. Alyasiri, Y.-N. Cheah, A. K. Abasi, and O. M. Al-Janabi, "Wrapper and hybrid feature selection methods using metaheuristic algorithms for English text classification: A systematic review," *IEEE Access*, vol. 10, pp. 39833–39852, 2022.
[22] A. Somasundaram and U. S. Reddy, "Data imbalance: Effects and solutions for classification of large and highly imbalanced data," in *Proc. Int. Conf. Res. Eng., Comput. Technol.*, 2016, pp. 1–16.
[23] K. Upadhyay, P. Kaur, and D. K. Verma, "Evaluating the performance of data level methods using KEEL tool to address class imbalance problem," *Arabian J. Sci. Eng.*, vol. 47, no. 8, pp. 9741–9754, Aug. 2022.
[24] S. Susan and A. Kumar, "The balancing trick: Optimized sampling of imbalanced datasets—A brief survey of the recent state of the art," *Eng. Rep.*, vol. 3, no. 4, Apr. 2021, Art. no. e12298.
[25] P. Kaur and A. Gosain, "Issues and challenges of class imbalance problem in classification," *Int. J. Inf. Technol.*, vol. 14, no. 1, pp. 539–545, Feb. 2022.
[26] P. Vuttipittayamongkol, E. Elyan, and A. Petrovski, "On the class overlap problem in imbalanced data classification," *Knowl.-Based Syst.*, vol. 212, Jan. 2021, Art. no. 106631.
[27] Y. Tang and J. Gao, "Improved classification for problem involving overlapping patterns," *IEICE Trans. Inf. Syst.*, vol. 90, no. 11, pp. 1787–1795, Nov. 2007.
[28] M. Denil and T. Trappenberg, "Overlap versus imbalance," in *Advances in Artificial Intelligence*, Ottawa, ONT, Canada. Springer, 2010, pp. 220–231.
[29] P. Peng and J. Wang, "Wear particle classification considering particle overlapping," *Wear*, vols. 422–423, pp. 119–127, Mar. 2019.
[30] R. Alejo, R. M. Valdovinos, V. García, and J. H. Pacheco-Sanchez, "A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios," *Pattern Recognit. Lett.*, vol. 34, no. 4, pp. 380–388, Mar. 2013.
[31] Y. Qu, H. Su, L. Guo, and J. Chu, "A novel SVM modeling approach for highly imbalanced and overlapping classification," *Intell. Data Anal.*, vol. 15, no. 3, pp. 319–341, May 2011.
[32] M. Ankerst, M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander, "Ordering points to identify the clustering structure," in *Proc. ACM SIGMOD*, vol. 99, 2008.
[33] I. Rish, "An empirical study of the naive Bayes classifier," in *Proc. IJCAI Workshop Empirical Methods Artif. Intell.*, 2001, vol. 3, no. 22, pp. 41–46.
[34] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, 1996, vol. 96, no. 34, pp. 226–231.
[35] K. M. Hasib, M. S. Iqbal, F. M. Shah, J. Al Mahmud, M. H. Popel, M. I. H. Showrov, S. Ahmed, and O. Rahman, "A survey of methods for managing the classification and solution of data imbalance problem," 2020, *arXiv:2012.11870*.
[36] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[37] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Proc. Int. Conf. Intell. Comput.* Springer, Aug. 2005, pp. 878–887.
[38] J. Mathew, M. Luo, C. K. Pang, and H. L. Chan, "Kernel-based smote for SVM classification of imbalanced datasets," in *Proc. 41st Annu. Conf. IEEE Ind. Electron. Soc.*, Nov. 2015, pp. 1127–1132.
[39] Y. Tang, Y.-Q. Zhang, N. V. Chawla, and S. Krasser, "SVMs modeling for highly imbalanced classification," *IEEE Trans. Syst., Man, Cybern., B*, vol. 39, no. 1, pp. 281–288, Feb. 2009.
[40] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Jun. 2008, pp. 1322–1328.
[41] J. Laaksonen and E. Oja, "Classification with learning k-nearest neighbors," in *Proc. Int. Conf. Neural Netw.*, vol. 3, 1996, pp. 1480–1483.
[42] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, "A support vector clustering method," in *Proc. 15th Int. Conf. Pattern Recognit.*, 2000, pp. 724–727.
[43] T. Kam Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, 1995, pp. 278–282.
[44] X. Zhang, Y. Li, R. Kotagiri, L. Wu, Z. Tari, and M. Cheriet, "KRNN: K rare-class nearest neighbour classification," *Pattern Recognit.*, vol. 62, pp. 33–44, Feb. 2017.
[45] H. Yu, C. Mu, C. Sun, W. Yang, X. Yang, and X. Zuo, "Support vector machine-based optimized decision threshold adjustment strategy for classifying imbalanced data," *Knowl.-Based Syst.*, vol. 76, pp. 67–78, Mar. 2015.
[46] B. Nikpour and H. Nezamabadi-Pour, "A memetic approach for training set selection in imbalanced data sets," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 11, pp. 3043–3070, Nov. 2019.
[47] Y. Zhu, Z. Wang, and D. Gao, "Gravitational fixed radius nearest neighbor for imbalanced problem," *Knowl.-Based Syst.*, vol. 90, pp. 224–238, Dec. 2015.
[48] S. Saryazdi, B. Nikpour, and H. Nezamabadi-Pour, "NPC: Neighbors' progressive competition algorithm for classification of imbalanced data sets," in *Proc. 3rd Iranian Conf. Intell. Syst. Signal Process. (ICSPIS)*, Dec. 2017, pp. 28–33.
[49] M. Tanveer, A. Sharma, and P. N. Suganthan, "Least squares KNN-based weighted multiclass twin SVM," *Neurocomputing*, vol. 459, pp. 454–464, Oct. 2021.
[50] B.-W. Yuan, X.-G. Luo, Z.-L. Zhang, Y. Yu, H.-W. Huo, T. Johannes, and X.-D. Zou, "A novel density-based adaptive k nearest neighbor method for dealing with overlapping problem in imbalanced datasets," *Neural Comput. Appl.*, vol. 33, no. 9, pp. 4457–4481, May 2021.
[51] Y. Li and X. Zhang, "Improving k nearest neighbor with exemplar generalization for imbalanced classification," in *Advances in Knowledge Discovery and Data Mining*, Shenzhen, China. Springer, 2011, pp. 321–332.
[52] X. Zhang and Y. Li, "A positive-biased nearest neighbour algorithm for imbalanced classification," in *Advances in Knowledge Discovery and Data Mining*. Springer, 2013, pp. 293–304.
[53] H. Dubey and V. Pudi, "Class based weighted k-nearest neighbor over imbalance dataset," in *Advances in Knowledge Discovery and Data Mining*. Springer, 2013, pp. 305–316.
[54] E. Kriminger, J. C. Príncipe, and C. Lakshminarayan, "Nearest neighbor distributions for imbalanced classification," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2012, pp. 1–5.
[55] Y. Song, J. Huang, D. Zhou, H. Zha, and C. L. Giles, "IKNN: Informative K-nearest neighbor pattern classification," in *Proc. Eur. Conf. Princ. Data Mining Knowl. Discovery*. Springer, 2007, pp. 248–264.
[56] R. Batuwita and V. Palade, "FSVM-CIL: Fuzzy support vector machines for class imbalance learning," *IEEE Trans. Fuzzy Syst.*, vol. 18, no. 3, pp. 558–571, Jun. 2010.
[57] H.-L. Dai, "Class imbalance learning via a fuzzy total margin based support vector machine," *Appl. Soft Comput.*, vol. 31, pp. 172–184, Jun. 2015.
[58] D. Gupta and B. Richhariya, "Entropy based fuzzy least squares twin support vector machine for class imbalance learning," *Int. J. Speech Technol.*, vol. 48, no. 11, pp. 4212–4231, Nov. 2018.
[59] T. Inkaya, "A density and connectivity based decision rule for pattern classification," *Exp. Syst. Appl.*, vol. 42, no. 2, pp. 906–912, Feb. 2015.

[60] J. Wang, P. Neskovic, and L. N. Cooper, "Improving nearest neighbor rule with a simple adaptive distance measure," *Pattern Recognit. Lett.*, vol. 28, no. 2, pp. 207–213, Jan. 2007.
[61] J. Alcalá-Fdez, L. Sánchez, S. García, M. J. del Jesus, S. Ventura, J. M. Garrell, J. Otero, C. Romero, J. Bacardit, V. M. Rivas, J. C. Fernández, and F. Herrera, "KEEL: A software tool to assess evolutionary algorithms for data mining problems," *Soft Comput.*, vol. 13, no. 3, pp. 307–318, Feb. 2009.
[62] *Keel: A Software Tool to Assess Evolutionary Algorithms for Data Mining Problems (Regression, Classification, Clustering, Pattern Mining and so on)*. Accessed: Mar. 7, 2022. [Online]. Available: https://sci2s.ugr.es/keel/datasets.php
[63] J. Huang, "Performance measures of machine learning," Univ. Western Ontario, Tech. Rep., 2008.

ZAHID AHMED received the B.Sc. degree in IT from Kuvempu University, Karnataka, the B.Tech. degree in computer science and engineering (CSE) from North-Eastern Hill University (NEHU), Shillong, Meghalaya, India, and the M.Tech. degree in CSE from Rajiv Gandhi Central University, Itanagar, Arunachal Pradesh. He is currently a Research Scholar with the Department of Information Technology (IT), NEHU. He has published several research papers in reputed journals and conferences. His area of research interests include imbalance classification and machine learning.

BIJU ISSAC (Senior Member, IEEE) received the Bachelor of Engineering (B.E.) degree in electronics and communication engineering, the Master of Computer Applications (M.C.A.) degree (Hons.), and the Ph.D. degree in networking and mobile communication. He joined Northumbria University, U.K., as an Academic Staff, in September 2018. He is currently an Associate Professor and the Head of the Subject (Networks and Cyber Security). He is the Director of the Academic Centre of Excellence in Cyber Security Research (ACE-CSR) and the Deputy Leader of the Cybersecurity and Networks (CyberNets) Research Group. He has published more than 100 research papers. He is a Chartered Engineer (C.Eng.) and a Senior Fellow of HEA and an EPSRC Associate Peer Review College Member.

SUFAL DAS received the bachelor's and master's degrees in engineering, in 2005 and 2008, respectively, and the Ph.D. degree from NEHU, in 2018. He has been an Assistant Professor with North-Eastern Hill University (NEHU), Shillong, India, since 2010. He has published several research papers in reputed journals and conferences. His area of research interests include big data analysis, data mining, and machine learning.