Complex & Intelligent Systems (2021) 7:2597–2615
https://doi.org/10.1007/s40747-021-00435-5

ORIGINAL ARTICLE

Scalable kernel-based SVM classification algorithm on imbalance air quality data for proficient healthcare

Shwet Ketu[1] · **Pramod Kumar Mishra**[1]

Received: 9 December 2020 / Accepted: 9 June 2021 / Published online: 29 June 2021

Abstract

In the last decade, we have seen drastic changes in the air pollution level, which has become a critical environmental issue. It should be handled carefully towards making the solutions for proficient healthcare. Reducing the impact of air pollution on human health is possible only if the data is correctly classified. In numerous classification problems, we are facing the class imbalance issue. Learning from imbalanced data is always a challenging task for researchers, and from time to time, possible solutions have been developed by researchers. In this paper, we are focused on dealing with the imbalanced class distribution in a way that the classification algorithm will not compromise its performance. The proposed algorithm is based on the concept of the adjusting kernel scaling (AKS) method to deal with the multi-class imbalanced dataset. The kernel function's selection has been evaluated with the help of weighting criteria and the chi-square test. All the experimental evaluation has been performed on sensor-based Indian Central Pollution Control Board (CPCB) dataset. The proposed algorithm with the highest accuracy of 99.66% wins the race among all the classification algorithms i.e. Adaboost (59.72%), Multi-Layer Perceptron (95.71%), GaussianNB (80.87%), and SVM (96.92). The results of the proposed algorithm are also better than the existing literature methods. It is also clear from these results that our proposed algorithm is efficient for dealing with class imbalance problems along with enhanced performance. Thus, accurate classification of air quality through our proposed algorithm will be useful for improving the existing preventive policies and will also help in enhancing the capabilities of effective emergency response in the worst pollution situation.

Keywords Air quality · Classification · Proficient healthcare · Scalable kernel-based SVM · Imbalance data

Introduction

In the machine learning paradigms, the classification of the new objects based on similar instances is one of the crucial tasks. The classification task becomes more complicated when one of the classes contains fewer instances than the other class [1]. The class imbalance problem is nothing but an unequal distribution of the data among the various classes. In the class imbalance problem, the majority of data samples belong to individual classes, and the rest of the data samples belong to the other classes. With respect to the binary class imbalance problem, one class contains the maximum number of data samples, and the other class contains only a few data samples [2]. The class which contains the maximum number of samples is said to be the majority class, and the class with the minimal number of samples is said to be the minority class [3, 4].

In the field of machine learning, it is one of the challenging tasks for classification algorithms to learn from imbalanced data. We are facing data imbalance issues in almost all the domains, or we can say that it is quite a common problem in all the fields. The areas which are facing these issues are the medical domain [5, 6], marketing doming, image classification [7], agriculture, big data domain [8–10], IoT [11–13], and so on [14–16]. Class imbalance is one of the critical issues in machine learning paradigms. If the classification algorithms are biased towards the majority class, then the accuracy of the classification algorithms will suffer much. Thus, if the new sample will come for classification, then it will be classified into the majority class because the classifier has lower prediction accuracy toward the minority

✉ Shwet Ketu
shwetiiita@gmail.com

Pramod Kumar Mishra
mishra@bhu.ac.in

[1] Department of Computer Science, Institute of Science, Banaras Hindu University, Varanasi, India

class. This situation is highly unappropriated and a severe matter of concern [17].

Nowadays, a drastic change in the air pollution level has been seen [18]. The pollution level in metropolitan cities is increasing, which not a very good sign for us. For making the environment healthier and comfortable, the air pollution level should be minimal. There are various liable factors that are making air polluted [19–22]. Some of them are directly, and some are indirectly participating in polluting the air. These pollutants are coming from various domains such as from industry, from transportation services, from daily traffic, from the thermal power plant, from various home appliances, garbage material from industries, hospitals and homes, and so on. The high level of air pollution can harm humans, animals, as well as botanical plants too [23]. Consequentially, a lot of new cases related to the respiratory system have been seen, which is the impact of bad air quality on human beings. It is also affecting crop quality and overall crop production. Thus, to reduce the effect of air pollution, we have to correctly classify the pollution level in real-time. From time to time, many researchers had contributed their approaches, which were accurate to some extent [24–28]. But due to the imbalanced nature of data, these models were not giving the correct prediction of the classes [29–32].

Building the classifier using the imbalanced datasets is one of the difficult tasks. In the classification task of imbalanced datasets, the minority class always suffers from the majority class because the classification model is biased with the majority class [33, 34]. As resultant, if any new sample comes for the classification, then it is classified in the majority class. This immanent need and gigantic interest motivate the researchers to deal with the class imbalance issues. From time to time, many researchers had given valuable solutions to deal with this class imbalance problem. These approaches were beneficial and capable of solving the problem to some extent by improving classifiers' performance. Most of the solutions were proposed for the binary class imbalance problems, which were not suitable for the multi-class imbalance problem. These limitations motivate us to deal with multi-class imbalance problems and also encourages us to give a possible contribution that will able to solve the multi-class imbalance problem. The contribution which we have worked on are:

- This solution is designed in a way that, which is well suited for both binary class and multi-class imbalance problems.
- The solution is based on algorithmic modification rather than data resampling at the processing phase.
- In our solution, the new kernel selection function has been proposed.

In this paper, the scalable kernel-based SVM (Support Vector Machine) classification algorithm has been proposed, which is capable of dealing with the multi-class data imbalance problem. First of all, the approximate hyperplane is gained using the standard SVM algorithm. After that, the weighting factor and parameter function for every support vector at each iteration is calculated. The values of these parameters are calculated using the Chi-square test. After that, the new kernel function or kernel transformation function is calculated. With the help of this kernel transformation function, the uneven class boundaries have been expanded, and the skewness of the data has been compensated. Therefore, the approximate hyperplane can be corrected by the proposed algorithm, and it can also resolve the performance degradation issue. In this study, we have also discussed the impact of air pollution on human health.

The rest of this paper has been arranged as follows. The related work has been drawn in "Related work". A brief discussion about the datasets, the working of the proposed algorithm with the mathematical foundation, and ten performance evaluation metrics have been briefly illustrated in "Materials and methods". The results of standard methods, existing literature methods and proposed classification algorithm have been presented in "Results". In "Discussion", the comprehensive conversation about the classification results and the effect of bad air quality on health have been discussed. Concluding remarks with the future scope have been drawn in "Conclusion".

Related work

Building the classifier using the imbalanced datasets is one of the difficult tasks. In the classification task of imbalanced datasets, the minority class always suffers from the majority class because the classification model is biased with the majority class [33, 34]. As resultant, if any new sample comes for the classification, then it is classified in the majority class. From time to time, many strategies have been made to overcome class imbalance issues. These proposed strategies are either work on the algorithm level or at the data level.

The data level approach is based on the resampling technique. Many classification algorithms such as SVM, naïve Bayes, C4.5, AdaBoost, and so on are using the resampling technique to deal with the data imbalance problem. The resampling task is consisting of two subtasks, i.e., under-sampling and over-sampling [35, 36]. The under-sampling technique is the process of filtering out the irrelevant sample from the dataset, and in the oversampling technique, we generate the new synthetic data. Two effective under-sampling methods had been proposed by Liu et al. [37], i.e., BalanceCascade and EasyEnsemble. In the BalanceCascade

Table 1 Classification algorithms to deal with data imbalance problem

Author	Approach	Objective	Algorithm	Result	Scope
Liu et al. [37]	Data level approach	Proposed two under-sampling method Balance-Cascade and Easy-Ensemble	EasyEnsemble and BalanceCascade	Deal with data imbalance	Used for data-level approach and for resolving data imbalance issues
Wang et al. [38]	Data level approach	An adaptive over-sampling approach has been proposed	Data density approach	Deals with data imbalance	Used for resolving data imbalance issue
Geo et al. [39]	Data level approach	Binary class over-sampling has been proposed	Using probabilistic methods	Deals with data imbalance	Used for resolving data imbalance issue
Batuwita and Palade [44]	Algorithmic level	Data imbalanced in the presence of noise	Fuzzy based SVM	Removing data imbalance	Classifier optimization
Cano et al. [45]	Algorithmic level	Proposed data imbalanced classifier	Gravitation weight-based	Removing data imbalance	Classifier optimization
Wu and Chang [46, 47]	Algorithmic level	Proposed boundary-based class boundary alignments	Improved SVM	Removing data imbalance	Classifier optimization
Oh et al. [48]	Algorithmic level	Proposed active sample election technique for data imbalance problem	Active sample election	Resolve data imbalance problem by improving performance	Increase the accuracy of the classifier
Liu et al. [49]	Algorithmic level	Proposed a sample selection technique	SVM	Increased the performance of the classifier	Increase the accuracy of the classifier
Fu and lee [51]	Algorithmic level	Proposed a certainty-based active learning algorithm	Machine learning	Resolve the data imbalanced and increase the performance	Active learning approach

technique, the sample that was correctly classified at each step was removed and did not participate in the further classification task. In the EasyEnsemble method, the majority class was divided into various subsets. These subsets had been used as input for the learner. The SMOTE stands for Synthetic Minority Over- Sampling Technique. It is one of the intelligent techniques which is based on an oversampling approach [36]. The oversampling in SMOTE is done by generating the syntactic samples for minority classes. The adaptive oversampling method had been proposed by Wang et al. [38], which was based on the data density approach. The binary class oversampling approach had been proposed by Geo et al. [39], which was based on the probability density function. Gu et al. [40] had discussed the data mining approaches on an imbalanced dataset.

The algorithmic level approaches are designed to bias the learning process to reduce the participation of the majority class and improve classifier performance. The solution for algorithmic level approaches has mainly consist of modification in algorithms, cost-sensitive learning, ensemble learning, and active learning.

The cost-sensitive learning approach is based on the concept of asymmetric cost assignment policy by minimizing the cost of misclassified samples. The cost minimization in a cost-sensitive approach is the process of penalizing the misclassified class with the penalty. But giving the desired penalty at every class level is an arduous task [41, 42]. Three cost-sensitive boosting algorithms for the classification of an imbalanced dataset in the AdaBoost framework had been introduced by Sun et al. [41]. The cost-sensitive SVM (support vector machine) had been proposed by Wang [43] to deal with the data imbalance problem.

To deal with the data imbalance problem, some researchers have done the modification in the algorithm level. The amendment in the algorithm level can be done at the classifier level by optimizing the classifier. The fuzzy-based SVM was proposed by Batuwita and Palade [44] to deal with imbalanced data in the presence of noise and outliers. Cano et al. [45] had proposed imbalanced data classification based on weighted gravitation. The adjusted boundary-based class boundary alignments with improved SVM performance had been proposed by Wu and Chang [46, 47].

The ensemble learning approach was designed to increase the accuracy of the classification algorithm. In this approach, several classifiers are used to train the model, and the decision output of these classifiers is incorporated into a single class. This final output is used for decision-making [3]. Bagging and boosting are the vital machine learning algorithm

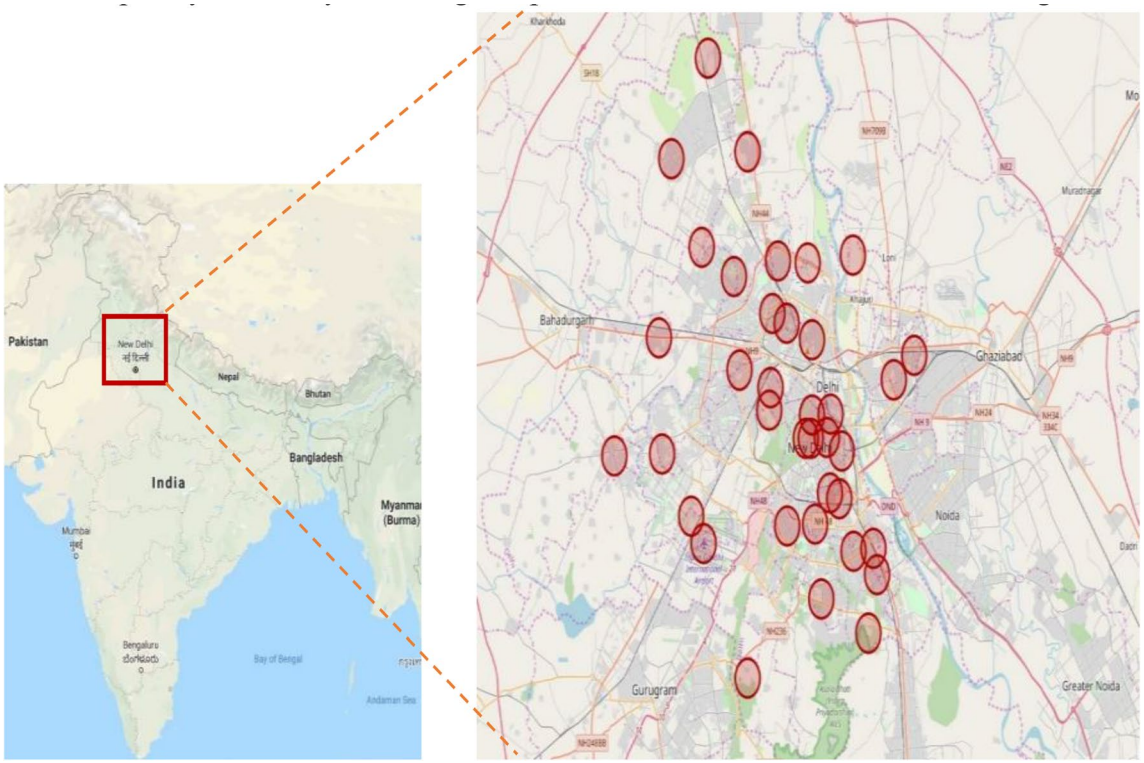

Fig. 1 Air quality data collection centers in the Delhi Region

in ensemble learning paradigms [3]. The active sample election technique was used by Oh et al. [48] to resolve the data imbalance problem. The sampling techniques (both under-sampling and over-sampling) had been integrated with the SVM to improve the classifier performance by Liu et al. [49].

The active learning approach is one of the exceptional cases of machine learning paradigms that have been used to label the new data sample points with the help of desired outputs by earning query interactively with a user [50]. The CBAL, which is a certainty-based active learning algorithm, was proposed by Fu and lee in 2013 [51] to solve the issue of data imbalance. Based on the various existing literature, the classification algorithms used to deal with the data imbalance problem have been briefly shown in Table 1.

Materials and methods

In this section, we will talk about the material and method which has been used in the experimental analysis. This section consists of three subsections, i.e., dataset illustration, proposed algorithm, and statistical measures. In the first subsection, the sensor-based CPCB dataset of Delhi has been discussed. In the second subsection, the proposed scalable kernel-based SVM classification algorithm has been discussed with its mathematical foundation. In the third subsection, a brief discussion about the performance evaluation metrics has been presented.

Data

For this study, we have taken the sensor-based CPCB data of Delhi city, which is the most polluted city in India. The reason behind taking this benchmark data is that the continuous monitoring of air quality with more than 200 base stations in approximately 20 states is maintained by the CPCB (Central Pollution Control Board). All the data from these stations are openly accessible from the website of CPCB. As far as Delhi is concerned, there are 37 base stations that are monitoring data continuously (24*7).

As we know, India comes second with respect to the total number of populations live after china [52, 53]. The massive growth in population is one of the key reasons for increasing pollution levels. Delhi is the capital and industrial hub of India; therefore, the population density of Delhi people is more than the respect of other cities. Resultant, the pollution caused by industrial waste and vehicles are the main reason for increasing the pollution level of Delhi [54, 55]. High discharge of various gasses, i.e., NO2, NH3, NO, CO2, O3,

and CO, with additional factors like wind direction, wind speed, temperature, and relative humidity make the air of Delhi heavily polluted and toxic. The toxic particles and other harmful particles are dissoluble in the air. Thus, living in such a polluted environment may cause some severe diseases. Even death is also possible in more severe cases. So, we have to take preventive measures to enhance the excellent quality of life by reducing the pollution levels for human well-being.

For the experimental analysis, the dataset from the Indian central pollution control board (CPCB) of capital Delhi has been taken [56]. The dataset has been extracted from various sensor-based devices. These sensor-based devices have been placed in multiple locations of Delhi and have been shown in Fig. 1. The figure has been plotted with the help of the longitude, and latitude of various data collection points falls of the Delhi region. The 37 data collection centers of Delhi have been plotted with the red circle in Fig. 1. We have taken the data from January 1, 2019, to October 1, 2020. The data has been recorded twenty-four times a day, which means, on an hourly basis. The CPCB air quality dataset is enriched with numerous liable features that can play an essential role in air quality classification tasks. These responsible features are PM10 (Concentration of Inhalable Particles), SO_2 (Sulfur Dioxide), PM2.5 (Fine Particulate Matter), O_3 (ozone), NOx (Nitrogen Oxide), NO_2 (Nitrogen Dioxide), NO (Nitrogen Monoxide), NH_3 (Ammonia), CO (Carbon Monoxide), AQI (Air Quality Index), WD (Wind Direction), C_6H_6 (Benzene), WS (Wind Speed), RH (Relative Humidity), SR (Solar Radiation), BP (Bar Pressure) and AT (Absolute Temperature). The dataset taken for the classification task contains 16 columns and 332,880 rows or 16 columns and 8760 rows at each base station (37 base stations are taken into consideration).

In this research work, the classification task has been performed on the CPCB air quality dataset, which contains various attributes. Only those attributes have been taken into consideration, which is responsible for making air pollution levels high. These attributes are the concentration of inhalable particles (PM10), sulfur dioxide (SO_2), fine particulate matter (PM2.5), ozone (O_3), nitrogen oxide (NOx), nitrogen dioxide (NO_2), nitrogen monoxide (NO), ammonia (NH3), carbon monoxide (CO), Air Quality Index (AQI) and, so on.

In Table 2, the various features of the dataset which are participating in the classification task have been presented. The various features have been explored with the help of several parameters, i.e., name of the variable with the abbreviation, nature of data, variable measuring unit, the period of data collection, variable type, and at last data extraction source.

Table 3 presents the various features of the dataset with the help of several parameters, such as the name of the variable with its mean values, measuring unit, standard

Table 2 Substantial features of the dataset a quick look

Variable	Variable Abbreviation	Nature of data	Measuring unit	Period of data collection	Variable type	Data source
Particulate Matter10	PM10	Real-Time	ug/m^3	01 Jan 2019 to 01 Oct 2020	Pollutant	CPCB
Sulfur Dioxide	SO2	Real-Time	ug/m^3	01 Jan 2019 to 01 Oct 2020	Pollutant	CPCB
Particulate Matter2.5	PM2.5	Real-Time	ug/m^3	01 Jan 2019 to 01 Oct 2020	Pollutant	CPCB
Ozone	O3	Real-Time	ug/m^3	01 Jan 2019 to 01 Oct 2020	Pollutant	CPCB
Nitrogen Oxide	NOx	Real-Time	Ppb	01 Jan 2019 to 01 Oct 2020	Pollutant	CPCB
Nitrogen Dioxide	NO2	Real-Time	ug/m^3	01 Jan 2019 to 01 Oct 2020	Pollutant	CPCB
Nitrogen Monoxide	NO	Real-Time	ug/m^3	01 Jan 2019 to 01 Oct 2020	Pollutant	CPCB
Ammonia	NH3	Real-Time	ug/m^3	01 Jan 2019 to 01 Oct 2020	Pollutant	CPCB
Carbon Monoxide	CO	Real-Time	ug/m^3	01 Jan 2019 to 01 Oct 2020	Pollutant	CPCB
Air Quality Index	AQI	Real-Time	ug/m^3	01 Jan 2019 to 01 Oct 2020	Pollution Level	CPCB

derivation, and actual and prescribed range of variables. These liable features have been used in the classification task.

Table 4 represents the data that has been come from the preprocessing and taken for the experimental analysis. This preprocessed data contains six classes, 270,596 samples, and ten attributes in each sample. The class-wise distribution of the dataset is 13,452, 47,910, 93,167, 55,045, 30,421, and 30,601 for class one to class six. The class imbalances ratio among the classes is 6.92.

Table 5 represents the description of the Air Quality Index (AQI), which contains the AQI range, appropriate AQI labeling, and class level. The labeling has been done into six parts according to the range from 0 to more than 400 [56]. The linking of the CPCB dataset with the AQI range is also established here.

Proposed methodology

The primary aim of the proposed algorithm is to deal with the data imbalance problem efficiently. The proposed algorithm is based on the concept of the adjusting kernel scaling method (AKS) [57] to deal with the multi-class imbalanced dataset. In this paper, we have proposed the SVM classification, which has been integrated with the adjusting kernel scaling method. In this section, a detailed discussion about the proposed algorithm has been presented.

Basic support vector machine algorithm (SVM)

Support Vector Machine (SVM) is a widely used and well-known machine learning algorithm for data classification. The SVM algorithm had been proposed by Vapnik et al. [58] in 1995. The primary aim of designing this algorithm was to map the input data into high dimensional space with the help of kernel function so that the classes can be linearly separable [58–60]. In the cased of the binary class problem, the maximum margin that can separate the hyperplanes is presented:

$$w.x + b = 0 \tag{1}$$

Based on the optimal pair (w_0, b_0), the decision function for SVM is represented by:

$$f(x) = \sum_{j \in SV} \lambda_j y_j \langle x.x_j \rangle + b \tag{2}$$

where, λ_j is support vector, x_j is data sample and $j = 1, 2, \ldots, C$.

Figure 2 shows the hyperplane with maximum separating margin and support vectors in the SVM algorithm paradigm.

For higher dimensional feature space, the value of $\langle x.x_j \rangle$ is replaced by the kernel function $K\langle x_j.x_i \rangle$ that is:

$$K\langle x_j.x_i \rangle = \langle x_j.x_i \rangle \tag{3}$$

Kernel function selection

In this section, the kernel function has been chosen from the standard SVM for approximately computing the boundaries' position. Initially, the dataset P is split into various samples which are $P^1, P^2, P^3, \ldots, P^j$ and after that, the kernel transformation function is applied that is defined in the below equation.

$$f(x) = \begin{cases} e^{-z_1 h(x)^2}, & if x \in P^1 \\ e^{-z_2 h(x)^2}, & if x \in P^2 \\ . \\ . \\ . \\ e^{-z_C h(x)^2}, & if x \in P^C \end{cases} \tag{4}$$

where, $h(x) = \sum_{j \in SV} \lambda_j y_j \langle x.x_j \rangle + b$ (where, λ_j is support vector), P^j is jth sample of the training set, the value of the parameter z_j is computed from the chi-square test (χ^2), which is explained in Sect. 2.2.2 and $j = 1,2, \ldots, C$.

Testing of Chi-square

The Chi-square test (χ^2) is one of the important statistical tests applied to sets of categorical features to determine the frequency distribution-based association among the categorical feature groups. In other words, we can say that it is used to evaluate the correlation among the groups. The significance of calculating the chi-square test is to establish the relationship among the samples of each category and parameter z_j. The mathematical formulation of evaluating the chi-square test (χ^2) is:

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e} \tag{5}$$

where, f_e and f_o are denoted as expected frequency and observed frequency, respectively.

Computing the weighting factor

The weighting factor is one of the important and difficult issues while dealing with the class imbalance problem. It is very difficult because assigning the appropriate weight for overcoming the class imbalance problem makes it complex. The simple way to deal with such problems is to give less weight to the majority class and more weight to the minority class by satisfying the weight condition $z_i \in (0,1)$.

The formulation of the weighting factor setting method has been used in the proposed algorithm to deal with the

Table 3 Variable description

Variable	Mean	Unit	Std. Dev	Prescribe range		Actual range	
				Min	Max	Min	Max
PM10	208.869	ug/m^3	154.392	0.00	100	0.14	1000
SO_2	106.398	ug/m^3	99.803	0.00	80	0.7	989.58
PM2.5	30.339	ug/m^3	55.716	0.00	60	0.01	499.1
O_3	51.994	ug/m^3	60.044	0.00	18	0.01	500
NOx	43.873	ppb	33.533	0.00	200	0.01	485.85
NO_2	35.515	ug/m^3	20.61	0.00	200	0.01	494.11
NO	14.821	ug/m^3	11.381	0.00	200	0.01	194.9
NH_3	1.362	ug/m^3	1.082	0.00	200	0.01	40.25
CO	41.407	ug/m^3	59.011	0.00	4	0.01	997
AQI	217.321	ug/m^3	152.63	0.00	100	8.85	1000

Table 4 Preprocessed dataset description

Dataset	CPCB (Central Pollution Control Board India)
Samples length	270,596
Number of Attributes	10
Number of Classes	6
Samples in each class	
Class 1	13,452
Class 2	47,910
Class 3	93,167
Class 4	55,045
Class 5	30,421
Class 6	30,601
Ratio of Imbalances	6.92

Table 5 Air quality description

AQI Range	Defined Labeling	Class belong to Dataset
0-50	Good	Class 1
50-100	Satisfactory	Class 2
100-200	Moderate	Class 3
200-300	Poor	Class 4
300-400	Very Poor	Class 5
400 +	Severe	Class 6

multi-class imbalance problem. In other words, we can say that the method that has been used to compensate for the uneven data distribution is defined as:

$$w_j = \frac{N/n_j}{\sum_{j=1}^{C} N/n_j} \tag{6}$$

where, N and C denote the training sample size and category size, respectively. nj indicates the sample size of every category with $j = 1, 2, \ldots, C$.

Computing the parameter z_j

Let P is the dataset, which includes the N number of samples with C categories. The value of the parameter z_j is calculated using Eqs. 2 and 3. The value of the chi-square (χ^2) in optimal distribution is,

$$\chi^2 = \sum_{j=1}^{C} \frac{(n_j - N/C)^2}{N/C} \tag{7}$$

where n_j = number of samples in jth category and $j = 1, 2, \ldots, C$

Let, $X_j = \frac{(n_j - N/C)^2}{N/C}$

Then,

$$\chi^2 = \sum_{j=1}^{C} X_j \tag{8}$$

So, the parameter z_j can be defined as

$$z_j = w_j \times \frac{X_j}{\chi^2} \tag{9}$$

From the Eq. (8) put the value of χ^2

$$z_j = w_j \times \frac{X_j}{\sum_{j=1}^{C} X_j} \tag{10}$$

Description of the proposed algorithm

The flow chart of the proposed algorithm has been shown in Fig. 3. First of all, the cleaning of the CPCB air quality dataset is performed, and after that, this proposed data is served to the classification algorithm for obtaining the initial partition. In the second step, we calculate the value of the weighting factor w_j and parameter z_j for every support vector in each iteration. The value of this parameter

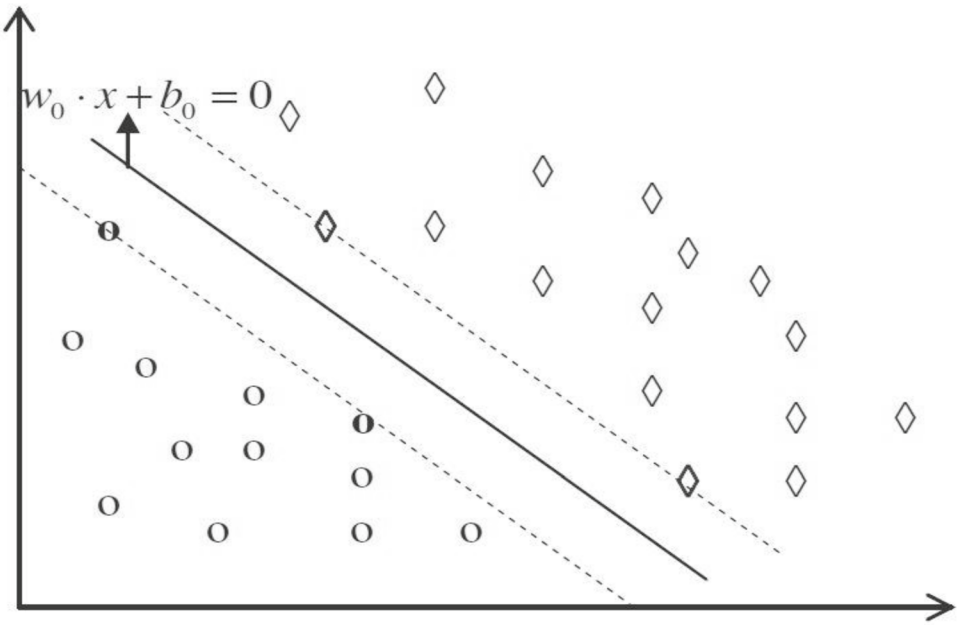

Fig. 2 Hyperplane with Support Vectors in the SVM Algorithm Paradigm

is calculated using the Chi-square test. In the next step, the kernel transformation function is calculated, and finally, the classification model is retrained using the new computed kernel matrix K_{mt}.

The algorithm for the proposed classification model contains 11 steps and these steps are described in Algorithm 1.

Algorithm 1. Procedure of the Proposed Algorithm

Step 1: START
Step 2: Initialization of SVM classifier with the training set X_{train} and kernel matrix $K = K_m$
Step 3: Based on training sample $x \in X_{train}$ the distance $h(x)$ is obtained with the initial partition of data $\{P^j, j = 1,2, \ldots, C\}$. ($C$ = No. of categories)
Step 4: $t \leftarrow 1$
Step 5: while $(t <= T)$ {
Step 6: Obtain the values of the parameters $z_j = w_j \times \frac{X_j}{\sum_{j=1}^{C} X_j}$ and $w_j = \frac{N/n_j}{\sum_{j=1}^{C} N/n_j}$
Step 7: Obtain the value of $f_{t-1}(x)$ for the training sample $x \in X_{train}$ by using Eqn. (17).
Step 8: The new kernel matrix K_{mt} is obtained by using the old kernel matrix (K_m) and $f_{t-1}(x)$
Step 9: Again, train the original SVM classifier with the training set X_{train} and kernel matrix K_{mt}.
Step 10: $i = i + 1$ }
Step 11: END

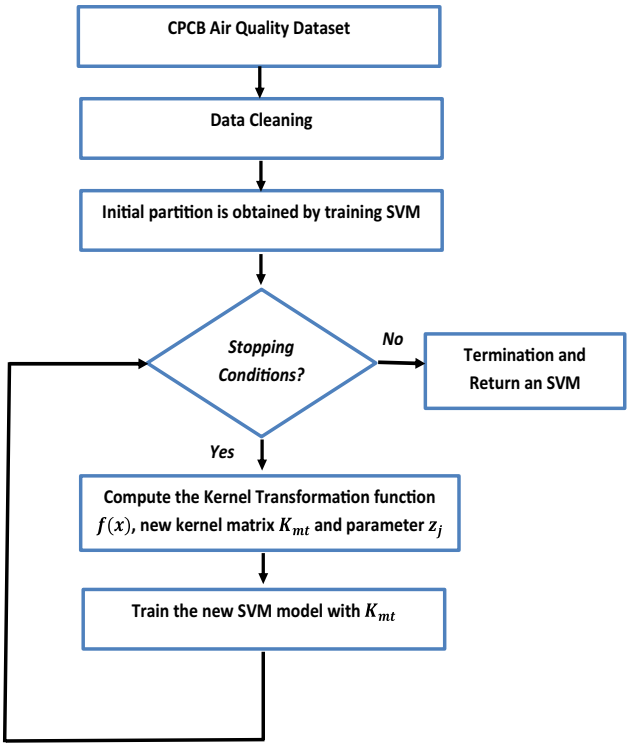

Fig. 3 Flow chart of the proposed algorithm

Statistical analysis

In this section, the various statistical measures used to evaluate the performance of the algorithms have been discussed. Statistical analysis is one of the essential tasks that help us pick the best algorithm based on their performance. In this paper, some statistical measures for evaluating the proposed algorithm and existing algorithms have been chosen to find the best algorithm among them. The statistical measures which have been taken into consideration are accuracy, precision, recall, f1-score, and TNR, NPV, FNR, FPR, FDR, FOR [61–64]. With the help of these ten evaluation measures, we can determine the appropriate algorithm that can perform the classification task more effectively and efficiently.

Accuracy

The accuracy with respect to the classification task is the percentage of instances that are correctly classified. In other words, we can say that accuracy is the percentage ratio of the correctly predicted class over the entire testing class. The accuracy formulation has been described in the below equation [61].

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \times 100\% \tag{11}$$

where TP, FP are the number of true positive and false positive respectively, and FN, TN represents the number of false negatives and true negatives, respectively.

Precision

The precision, with respect to the classification task, is used to quantify the number of predicted positive classes that are actually falling under the positive class. In other words, we can say that precision is the ratio of the true positive class over the total number of a truly positive and false-positive class. The precision formulation has been described in the below equation [61, 63].

$$Precision = \frac{(TP)}{(TP + FP)} \tag{12}$$

where, TPandFP are the numbers of true positive and false positive, respectively.

Recall

The recall, with respect to the classification task, is used to quantifies the number of predicted positive class actually falls out of all positive instances in the dataset. In other words, we can say that recall is the ratio of the true positive class over the total number of a truly positive and false negative class. The recall formulation has been described in the below equation [63].

$$Recall = \frac{(TP)}{(TP + FN)} \tag{13}$$

where, TPandFN are the numbers of true positive and false negative, respectively.

F1-score

The F1- score is also known as F Measure or F Score. The F1-score, with respect to the classification task, is used to quantifies the balance among the recall and precision. In other words, we can say that F1-score is the twice product of recall and precision over the summation of recall and precision. The f1-score formulation has been described in the below equation [62].

$$f1 = \frac{2 \times (precision \times recall)}{precision + recall} \tag{14}$$

True negative rate (TNR)

The TNR, with respect to the classification task, is used to quantifies the specificity or true negative rate. In other words, we can say that TNR is the ratio of the true negative class over the total number of a truly negative and false positive class. The TNR formulation has been described in the below equation [61, 64].

$$TNR = \frac{(TN)}{(TN + FP)} \tag{15}$$

where, TN and FP are the numbers of true negative and false positive, respectively.

Negative predictive value (NPV)

The NPV, with respect to the classification task, is used to quantifies the ratio of negative predictive value. In other words, we can say that NPV is the ratio of the true positive class over the total number of a truly positive and false negative class. The NPV formulation has been described in the below equation [61, 64].

$$NPV = \frac{(TN)}{(TN + FN)} \tag{16}$$

where, TN and FN are the numbers of true and false negative, respectively.

False negative rate (FNR)

The FNR, with respect to the classification task, is used to quantifies the miss rate. In other words, we can say that FNR is the ratio of the false-negative class over the total number of a truly positive and false negative class. The FNR formulation has been described in the below equation [61, 64].

$$FNR = \frac{(FN)}{(FN + TP)} \tag{17}$$

where, TP and FN are the numbers of true positive and false negative, respectively.

False positive rate (FPR)

The FPR, with respect to the classification task, is used to quantifies the fall-out rate. In other words, we can say that FPR is the ratio of the false positive class over the total number of a truly negative and false positive class. The FPR formulation has been described in the below equation [61, 64].

$$FPR = \frac{(FP)}{(FP + TN)} \tag{18}$$

where, FP and TN are the numbers of false-positive and true negative, respectively.

False discovery rate (FDR)

The FDR is the ratio of the false positive class over the total number of a truly positive and false-positive class. The FDR formulation has been described in the below equation [61, 64].

$$FDR = \frac{(FP)}{(FP + TP)} \tag{19}$$

where, FP an dTP are the numbers of false and true negative, respectively.

False omission rate (FOR)

The FOR is the ratio of the false-negative class over the total number of a truly negative and false negative class. The FOR formulation has been described in the below equation [61, 64].

$$FOR = \frac{(FN)}{(FN + TN)} \tag{20}$$

where, TN an dFN are the numbers of true and false negative, respectively.

Results

In this section, we will talk about the classification result based on classification algorithm, i.e., Ada Boost Algorithm (ADB) [65–67], Multilayer Perceptron Algorithm (MLP) [68–70], Gaussian NB Algorithm (GNB) [71–73], standard Support Vector Machine Algorithm (SVM) [58–60], existing literature methods and proposed scalable-kernel based SVM algorithm.

Model comparison

Identifying the best classification model capable of dealing with class imbalance problems is one of the complex tasks. The CPCB air quality dataset has been taken for the experimental analysis. In Fig. 4, the x-axis denotes the various classes, and the y-axis indicates the number of data samples in the multiple classes. From Fig. 4, it is clear that our dataset contains uneven class distribution, or we can say that it is imbalanced. Therefore, it becomes more challenging to handle such a situation by the traditional classification models. The class-wise distributions of the dataset based on sample size are: the first class consists of 13,452 samples, the second one contains 47,910 samples, the third one has 93,167

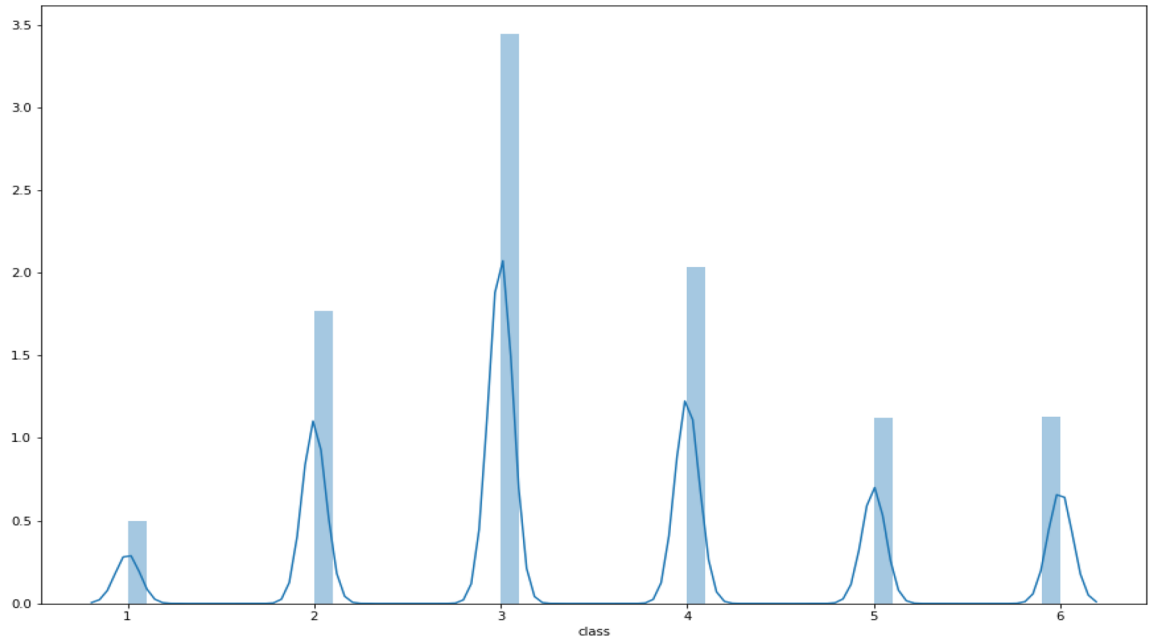

Fig. 4 Class wise distribution of CPCB dataset

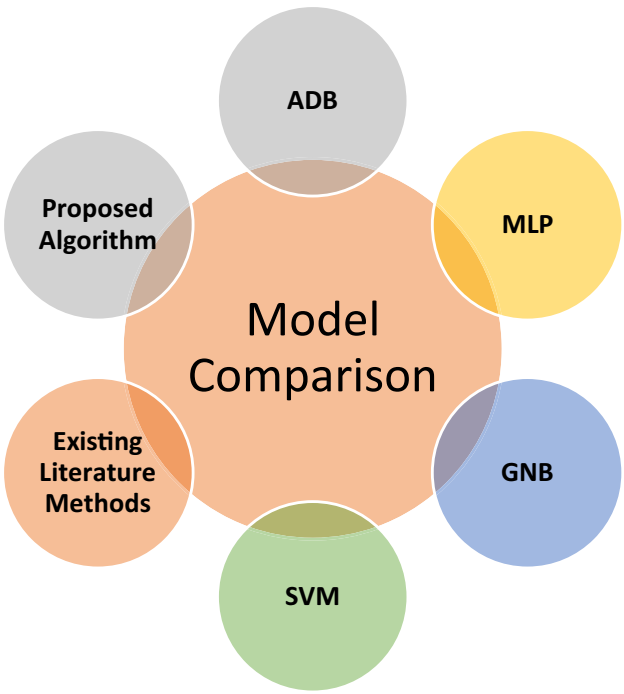

Fig. 5 Classification models for experimental evaluation

samples, the fourth class has 55,045 samples, the fifth class contains 30,421 samples, and the last class contains 30,601 samples. The dataset also has a 6.92 class imbalance ratio.

The primary aim of this research work to find out the best classification model which can deal with the class imbalance problem. From time to time, many researchers had given valuable solutions to deal with this class imbalance problem. Most of the solutions were proposed for the binary class imbalance problems and which did not find suitable for the multi-class imbalance problem. These limitations motivate us to modify the algorithm that can efficiently deal with multi-class and binary class imbalance problems without compromising the algorithm's performance. This classification will also be helpful for making the possible solution toward proficient healthcare.

For the experimental evaluation, the four well-established traditional classification algorithms and existing literature methods with our proposed algorithm have been taken. Our proposed algorithm has been compared with other algorithms to determine suitability, correctness, and efficiency. The ten performance validation measures have measured the performance of all the classification algorithms. The tenfold cross-validation policy has been used.

Figure 5 shows an overview of the classification algorithms, which have been used in the classification task. The four classification algorithms and existing literature methods have been compared with our proposed algorithm to determine the performance of the proposed classifier. The algorithms which have been used in the classification task are ADB (Ada Boost Algorithm), MLP (Multilayer Perceptron Algorithm), GNB (Gaussian NB Algorithm), standard SVM (Support Vector Machine Algorithm), existing literature methods, and proposed scalable-kernel based SVM algorithm.

Table 6 Performance evaluation of classification algorithms I

Classification results for real-time sensor generated air quality index (AQI) dataset										
Classifier name	Real-time sensor generated air quality index (AQI) dataset representation									
	Precision	Recall	F1-Score	TNR	NPV	FNR	FPR	FDR	FOR	Accuracy
Ada Boost classifier	0.48	0.60	0.46	0.59	0.59	0.41	0.08	0.41	0.41	59.72
MLP classifier	0.96	0.97	0.96	0.95	0.95	0.03	0.01	0.03	0.05	95.71
Gaussian NB	0.81	0.81	0.81	0.80	0.80	0.19	0.03	0.19	0.2	80.87
SVM classifier	0.97	0.97	0.97	0.96	0.96	0.03	0.01	0.03	0.04	96.92
Proposed algorithm	1.00	1.00	1.00	0.99	0.99	0.002	0.001	0.002	0.01	99.66

*TNR- True Negative Rate, NPV- Negative Predictive Value, FNR- False Negative Rate, FPR- False Positive Rate, FDR- False Discovery Rate, FOR- False Omission Rate

Table 7 Performance evaluation of existing literature methods vs proposed classification algorithm

Model used	Accuracy (%)
Existing literature methods	
Cost-sensitive Boosting[41]	90.52
Cost-sensitive SVM [43]	95.01
Fuzzy based SVM [44]	97.19
Improved SVM [46]	97.51
Impoved SVM [49]	96.90
Proposed Model	
Scalable kernel based SVM	**99.66**

Performance evaluation of classification algorithms

The performance evaluation of the classification algorithm has been divided into two parts. In the first part, the CPCB air quality dataset of the whole Delhi region has been taken, which has been come from the 37 distributed base stations. All the data has been served as a single file to perform the classification task. The classification result of all algorithms has been evaluated in the form of precision, recall, F1 score, TNR, NPV, FNR, FPR, FDR, FOR, and accuracy. The validation of the classification algorithm has been performed based on the classification accuracy. As we know, our dataset contains imbalanced class distribution that may affect the classification algorithms' performance. All standard models perform well except Ada Boost Classifier (ADB). The ADB classifier achieves the lowest accuracy of 59.72 among all the classifiers. The standard SVM classifier, MLP classifier, and Gaussian NB perform quite well in imbalanced class distribution. But if we compare it to our proposed SVM classifier, then these classifiers are lost in the battle. Our proposed algorithm wins the battle with the highest accuracy of 99.66 among all the other models. The detailed analysis of the classification results has been shown in Table 6.

In the second part of the performance evaluation, we have taken the CPCB dataset, which is coming from 37 places in Delhi. The proposed algorithm achieves the highest accuracy of 99.66% among the existing literature methods. It is also efficient for dealing with class imbalance problems without compromising performance. Performance evaluation of existing literature method Vs proposed classification algorithm has been presented in Table 7.

In the second part of the performance evaluation, we have taken the individual data of each base station of CPCB, which is plotted at 37 places in Delhi. The 37 data files have been used as input datasets for performing the classification task with the help of various classification algorithms. The details about the acronyms used in Table 8 have been defined in Appendix 1. Our proposed algorithm has performed exceptionally well in this rigorous analysis for all the datasets lying from A1 to A37. Our proposed algorithm achieves the highest average accuracy of 99.72 (average of A1 to A 37) among all the algorithms. It is also efficient for dealing with class imbalance problems without compromising performance. The detailed analysis of the results has been shown in Table 8.

Discussion

Numerous associated factors exist which may play a crucial role in affecting air quality. Some factors directly and some indirectly participate in polluting the air. Those pollutants which are air soluble are hazardous to human health. The poor diffusion condition is one of the crucial factors which play a vital role in increasing the level of pollutant. The drive of air partials from the high concentration space to the low concentration space is known as diffusion. Before performing the classification task, the preprocessing of the dataset is performed. Preprocessing is a process of dropping missing values and the unusual object from the datasets. The dataset is consist of numerous liable features such as PM10 (Concentration of Inhalable Particles), SO2 (Sulfur Dioxide), PM2.5 (Fine Particulate Matter), O3 (ozone), NOx (Nitrogen Oxide), NO2 (Nitrogen Dioxide), NO (Nitrogen Monoxide), NH3 (Ammonia), CO (Carbon Monoxide), AQI (Air Quality Index), WD (Wind Direction), C6H6 (Benzene), WS (Wind

Table 8 Performance evaluation of classification algorithms II

Data collected from	Classifiers				
	ADB	MLP	GNB	SVM	Proposed algorithm
A1	67.00	86.40	83.13	94.81	**99.67**
A2	53.14	74.77	82.88	94.47	**99.51**
A3	68.92	75.45	81.06	95.12	**99.65**
A4	66.10	81.39	84.80	95.07	**99.67**
A5	84.78	90.24	82.71	94.96	**99.52**
A6	67.44	83.42	86.00	94.29	**99.56**
A7	68.54	90.26	80.92	93.27	**99.59**
A8	67.67	90.98	81.44	95.83	**99.95**
A9	68.12	84.92	85.05	95.68	**99.73**
A10	73.94	91.50	85.13	97.53	**99.86**
A11	60.74	86.13	83.20	95.41	**99.67**
A12	85.69	90.33	82.55	96.49	**99.79**
A13	67.83	87.26	78.62	96.50	**99.81**
A14	97.58	65.50	82.86	95.20	**99.77**
A15	63.94	85.63	83.23	93.68	**99.25**
A16	66.93	85.48	82.03	96.55	**99.41**
A17	63.91	77.76	81.46	95.94	**99.64**
A18	73.02	91.95	81.62	96.54	**99.95**
A19	68.22	89.95	84.67	97.40	**99.45**
A20	70.82	87.94	80.63	94.51	**100**
A21	69.35	85.57	84.79	95.99	**99.78**
A22	75.54	74.08	82.70	95.60	**99.95**
A23	81.24	91.02	82.18	94.81	**99.81**
A24	73.75	90.84	84.18	94.94	**99.72**
A25	69.88	77.36	83.29	96.56	**99.85**
A26	92.20	83.16	79.97	94.83	**99.53**
A27	66.77	79.01	82.08	96.50	**99.82**
A28	65.78	84.26	83.95	95.08	**99.86**
A29	72.48	92.01	83.52	96.87	**99.87**
A30	65.24	87.00	84.39	94.90	**99.5**
A31	64.33	91.36	80.58	94.63	**99.71**
A32	69.91	85.12	80.55	92.09	**99.67**
A33	88.91	92.78	83.11	96.48	**99.78**
A34	72.87	91.05	80.34	95.72	**99.76**
A35	63.84	91.09	83.50	95.12	**99.91**
A36	82.97	85.59	85.44	96.92	**99.79**
A37	79.14	71.17	84.12	96.63	**99.76**
Total accuracy	**71.85**	**85.13**	**82.78**	**95.48**	**99.72**

Speed), RH (Relative Humidity), SR (Solar Radiation), BP (Bar Pressure) and AT (Absolute Temperature). The correlation on the preprocessed data is calculated to find out the relation between the class and the liable factors.

In Fig. 6, the relationship between class and respondent factors has been shown. With the help of correlations, we can easily find which responsive factors are highly correlated with the class.

Performance evaluation of classification algorithms

For the experimental analysis, the dataset from the Indian central pollution control board (CPCB) of the capital Delhi has been taken. The data from January 1, 2019, to October 1, 2020, has been used for training and testing purposes. The tenfold cross-validation policy has been used. Cross-validation is a technique to assess models by partitioning the given data sample into the training and testing sets. The training set is used to train the model whereas the testing set to evaluate the model. In k-fold cross-validation, the given data sample is randomly divided into the k subsamples of equal size. Where the k-1 subsample is used for training the model and a single subsample is used for validation purposes. This cross-validation technique is repeated up to k times (k- folds) and each subsample is used exactly once for validation purposes. The single estimation is produced by averaging all the result fall under k-fold. The algorithms which have been used in the classification task are ADB (Ada Boost Algorithm), MLP (Multilayer Perceptron Algorithm), GNB (Gaussian NB Algorithm), standard SVM (Support Vector Machine Algorithm), existing literature methods, and proposed scalable kernel-based SVM algorithm.

In Fig. 7, the experimental results, (i.e. statistical measures based and existing literature methods versus proposed algorithm) of the various classification algorithms on the CPCB dataset of the whole Delhi region have been presented. From the figure, it is clear that our proposed algorithm with the highest accuracy of 99.66 wins the race among all the classification algorithms and existing literature methods. The result of the proposed algorithm is also better than the traditional SVM algorithm. So, it is also clear from the results that our proposed algorithm is efficient for dealing with class imbalance problems without compromise the performance of the algorithm.

In Fig. 8, the accuracy-based classification results of the various classification algorithms on the CPCB dataset, specifically A1, A10, A20, A30, and A37 of the Delhi region have been plotted using a bar graph. From the figure, it is clear that our proposed algorithm achieves the highest accuracy throughout the areas and wins the race among the classification algorithms. The results of the proposed algorithm are also better than the traditional SVM algorithm. Thus, it is also clear from the results that our proposed algorithm is efficient for dealing with class imbalance problems along with enhanced performance.

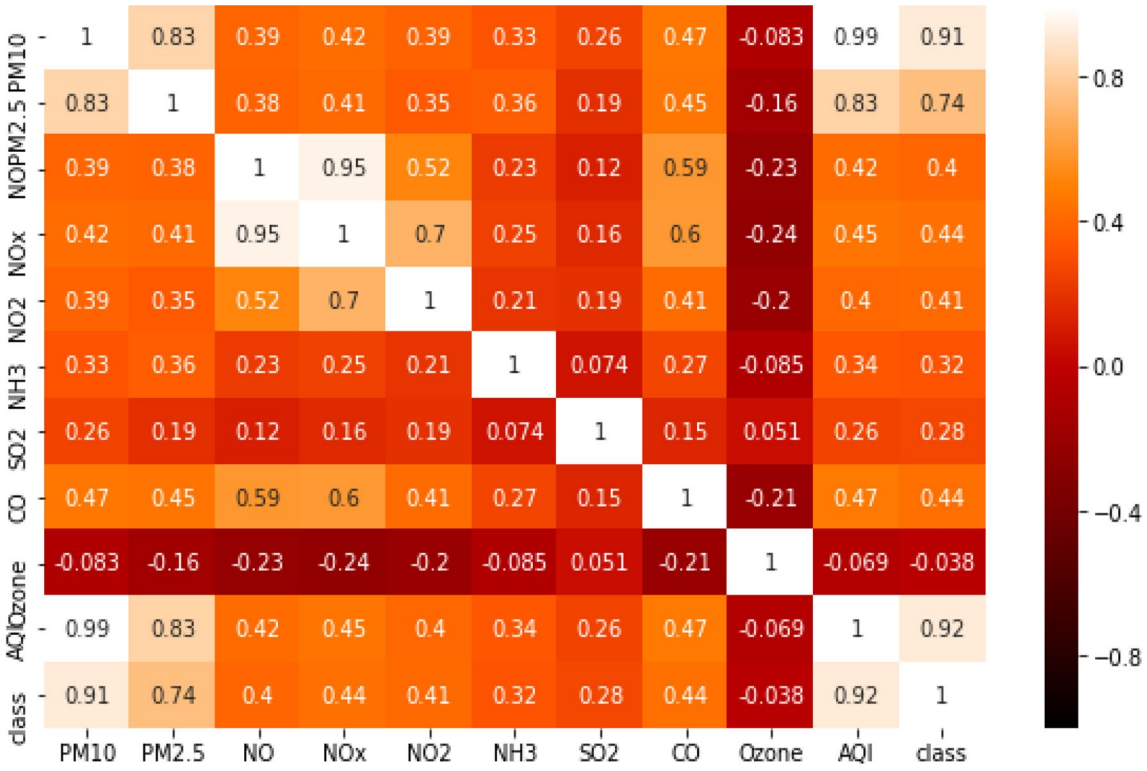

Fig. 6 Correlation coefficients of liable factor

Effect on healthcare

Bad air quality can impact individuals' health and quality of life. The impact of bad air quality may cause problems from minor to severe. It may affect individuals' cardiovascular or circulatory system, respiratory system, excretory system (kidney or urinary), nervous system, endocrine system, circulatory system, digestive system, lymphatic system, integumentary system (skin), and ophthalmic system.

Table 9 shows the AQI range with associated labeling, and the impact of various air levels on health has been shown [56]. The AQI level is divided into the six-range starting from 0–50 and end at greater than 400.

The consequence of high AQI levels on individuals' health has been described in Table 10. The various effects of high AQI levels are divided into three subparts, i.e., short-term impact, long-term impact, and severe impact. It may cause severe problems for those people who are suffering from respiratory diseases. Such people require intensive care, and precaution must be taken to minimize its impact on their health [74–77].

Conclusion

In numerous classification problems, we are facing the class imbalance issue. This research is focused on dealing with the imbalanced class distribution so that the classification algorithm will not compromise its performance. The proposed algorithm is based on the concept of the adjusting kernel scaling (AKS) method to deal with the multi-class imbalanced dataset. The scalable kernel-based SVM classification algorithm has been proposed and presented in this study. In the proposed algorithm, the kernel function's selection has been evaluated based on the weighting criteria and chi-square test. Using this kernel transformation function, the uneven class boundaries have been expanded, and the skewness of the data has been compensated. For experimental evaluation, we have taken the accuracy-based classification results of the various classification algorithms on the CPCB dataset of Delhi to find and evaluate the performance of our proposed algorithm over the other classification algorithms. Our proposed algorithm with the highest accuracy 99.66% wins the race among all classification algorithms,

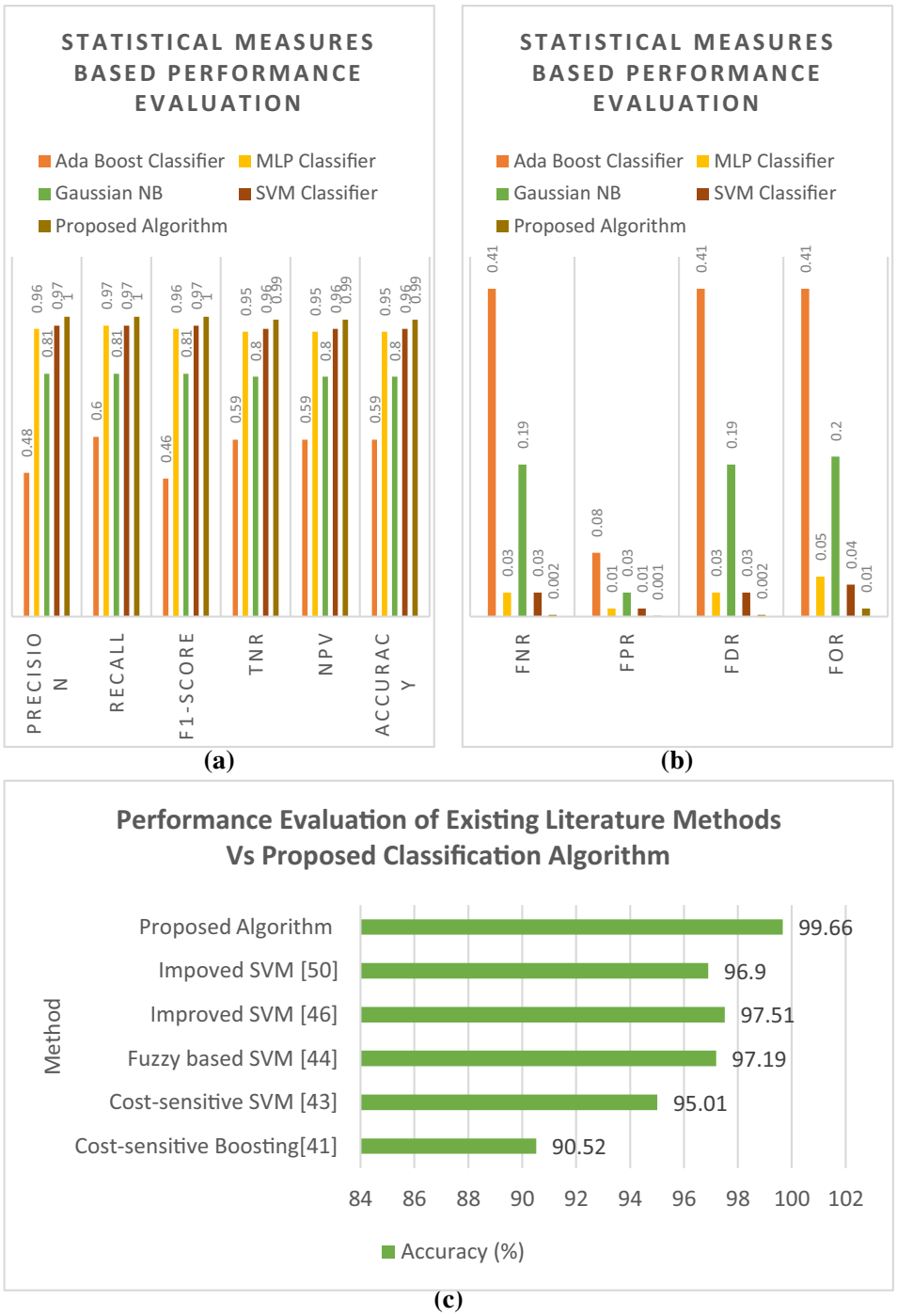

Fig. 7 Result of the classification algorithms. **a** Statistical Measures based I. **b** Statistical Measures based II. **c** Existing Literature Methods Vs Proposed Algorithm

and the result of the proposed algorithm is even better than the traditional SVM algorithm. The results of the proposed algorithm are also better than the existing literature methods. It is also clear from these results that our proposed algorithm is efficient in dealing with the class imbalance and enhanced performance. In this study, we have also discussed the effect of air pollution on human health, which is possible only if the data are correctly classified. Thus, accurate air quality

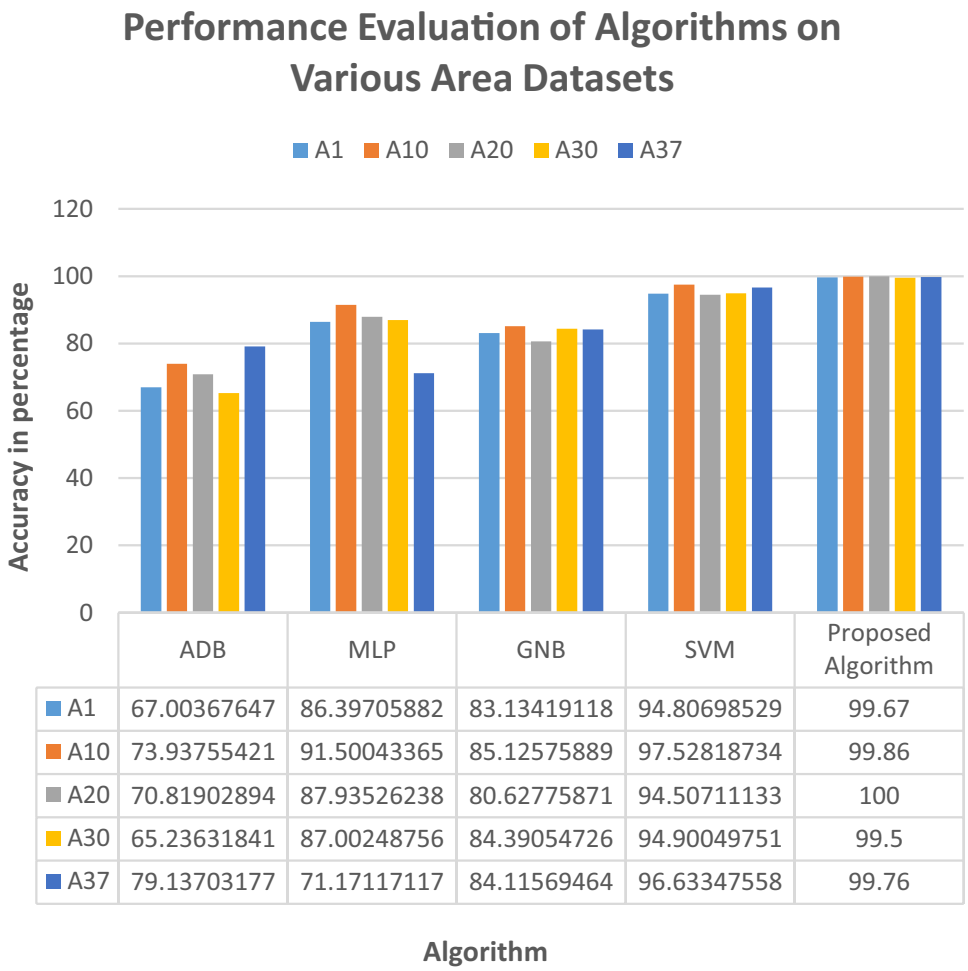

Fig. 8 Accuracy-based results of the classification algorithms II

	ADB	MLP	GNB	SVM	Proposed Algorithm
A1	67.00367647	86.39705882	83.13419118	94.80698529	99.67
A10	73.93755421	91.50043365	85.12575889	97.52818734	99.86
A20	70.81902894	87.93526238	80.62775871	94.50711133	100
A30	65.23631841	87.00248756	84.39054726	94.90049751	99.5
A37	79.13703177	71.17117117	84.11569464	96.63347558	99.76

Table 9 Air quality index range with possible health impact

AQI Range	Labeling	Impact of health
0-50	Good	Minor Impact
50-100	Satisfactory	Discomfort for sensitive people such as minor breathing problem
100-200	Moderate	May cause breathing problems for the people who are suffering from diseases related to the lungs and heart.
200-300	Poor	May have Breathing problem in most of the people live in this situation.
300-400	Very Poor	May have Respiratory sickness in most of the people live in this situation
400 +	Severe	May have a Serious problem with Healthy and ill people.

classification through our proposed algorithm would be useful for improving the existing preventive policies and would also help enhance the capabilities of effective emergency response in the event of the worst pollution.

In the future, this algorithm will be compared with the recent variation of SVM. The proposed algorithm will be tested on other datasets, and we will try to improve its computational methods as well.

Table 10 The effect of high air quality index level on person's health

Pollutants	AQI	
Effect on health	Short term	1. Serious cardiovascular illness 2. Serious respiratory illness 3. Cause more strain on lungs and heart 4. Damaged respiratory system cells
	Long term	1. Faster aging of the lungs 2. Reduction of lung capacity 3. Reduction in lungs functionality 4. Bronchitis 5. Asthma 6. Possibly cancer 7. Emphysema 8. Shorter life span
	Severe health problems for	1. The person suffering from heart disease 2. The person suffering from congestive heart failure 3. The person suffering from coronary artery syndrome 4. The person suffering from asthma 5. The person suffering from Emphysema 6. The person suffering from COPD (Chronic Obstructive Pulmonary Disease) 7. Women with Pregnancy 8. Outdoor labors 9. Old age people and children below 14 years of age 10. Sportsperson who exercise strongly outdoors

Appendix 1

See Table 11.

Table 11 List of abbreviations

S. no	Acronyms	Full form
1	A1	Alipur, Delhi – DPCC
2	A2	Anand Vihar, Delhi – DPCC
3	A3	Ashok Vihar, Delhi – DPCC
4	A4	Aya Nagar, Delhi – IMD
5	A5	Bawana, Delhi – DPCC
6	A6	Burari Crossing, Delhi – IMD
7	A7	CRRI Mathura Road, Delhi – IMD
8	A8	Dr. Karni Singh Shooting Range, Delhi—DPCC
9	A9	DTU, Delhi – CPCB
10	A10	Dwarka-Sector 8, Delhi – DPCC
11	A11	IGI Airport (T3), Delhi – IMD
12	A12	IHBAS, Dilshad Garden, Delhi—CPCB
13	A13	ITO, Delhi – CPCB
14	A14	Jahangirpuri, Delhi – DPCC
15	A15	Jawaharlal Nehru Stadium, Delhi—DPCC
16	A16	Lodhi Road, Delhi – IMD
17	A17	Major Dhyan Chand National Stadium, Delhi—DPCC
18	A18	Mandir Marg, Delhi – DPCC
19	A19	Mundka, Delhi – DPCC
20	A20	Najafgarh, Delhi – DPCC
21	A21	Narela, Delhi – DPCC
22	A22	Nehru Nagar, Delhi – DPCC
23	A23	North Campus, DU, Delhi – IMD
24	A24	NSIT Dwarka, Delhi – CPCB
25	A25	Okhla Phase-2, Delhi – DPCC
26	A26	Patparganj, Delhi – DPCC
27	A27	Punjabi Bagh, Delhi – DPCC
28	A28	Pusa, Delhi – DPCC
29	A29	Pusa, Delhi – IMD
30	A30	R K Puram, Delhi – DPCC
31	A31	Rohini, Delhi – DPCC
32	A32	Shadipur, Delhi – CPCB
33	A33	Sirifort, Delhi – CPCB
34	A34	Sonia Vihar, Delhi – DPCC
35	A35	Sri Aurobindo Marg, Delhi – DPCC
36	A36	Vivek Vihar, Delhi – DPCC
37	A37	Wazirpur, Delhi – DPCC

Declarations

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

References

1. Menardi G, Torelli N (2014) Training and assessing classification rules with imbalanced data. Data Min Knowl Disc 28(1):92–122
2. Japkowicz N, Stephen S (2002) The class imbalance problem: a systematic study. Intell Data Anal 6(5):429–449
3. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F (2011) A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans Syst Man Cybern Part C (Appl Rev) 42(4):463–484
4. Wang S, Yao X (2012) Multi-class imbalance problems: analysis and potential solutions. IEEE Trans Syst Man Cybern Part B (Cybern) 42(4):1119–1130
5. Ketu S, Mishra PK (2021) Hybrid classification model for eye state detection using electroencephalogram signals. Cognit Neurodyn 1–18
6. Ketu S, Mishra PK (2020). A hybrid deep learning model for COVID-19 prediction and current status of clinical trials worldwide. Comput Mater Contin 66(2)
7. Tali RV, Borra S, Mahmud M (2021) Detection and classification of leukocytes in blood smear images: state of the art and challenges. Int J Ambient Comput Intell (IJACI) 12(2):111–139
8. Ketu S, Agarwal S (2015) Performance enhancement of distributed K-Means clustering for big Data analytics through in-memory computation. In: 2015 Eighth international conference on contemporary computing (IC3), IEEE, pp 318–324
9. Ketu S, Prasad BR, Agarwal S (2015) Effect of corpus size selection on performance of map-reduce based distributed k-means for big textual data clustering. In Proceedings of the sixth international conference on computer and communication technology 2015, pp 256–260
10. Ketu S, Kumar Mishra P, Agarwal S (2020). Performance analysis of distributed computing frameworks for big data analytics: hadoop vs spark. Comput Sistemas 24(2)
11. Ketu S, Mishra PK (2020) Performance analysis of machine learning algorithms for IoT-based human activity recognition. In Advances in electrical and computer technologies, pp 579–591, Springer, Singapore
12. Ketu S, Mishra PK (2021) Enhanced Gaussian process regression-based forecasting model for COVID-19 outbreak and significance of IoT for its detection. Appl Intell 51(3):1492–1512
13. Ketu S, Mishra PK (2021) Cloud, fog and mist computing in IoT: an indication of emerging opportunities. IETE Tech Rev, pp 1–12
14. Chawla NV, Japkowicz N, Kotcz A (2004) Special issue on learning from imbalanced data sets. ACM SIGKDD Explor Newsl 6(1):1–6
15. Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD (2008) Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance. Neural Netw 21(2–3):427–436
16. Kubat M, Holte RC, Matwin S (1998) Machine learning for the detection of oil spills in satellite radar images. Mach Learn 30(2–3):195–215
17. Daskalaki S, Kopanas I, Avouris N (2006) Evaluation of classifiers for an uneven class distribution problem. Appl Artif Intell 20(5):381–417
18. Vitousek PM (1994) Beyond global warming: ecology and global change. Ecology 75(7):1861–1876
19. Yilmaz O, Kara BY, Yetis U (2017) Hazardous waste management system design under population and environmental impact considerations. J Environ Manag 203:720–731
20. De Vito S, Piga M, Martinotto L, Di Francia G (2009) CO, NO2 and NOx urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization. Sens Actuators B Chem 143(1):182–191
21. Northey SA, Mudd GM, Werner TT (2018) Unresolved complexity in assessments of mineral resource depletion and availability. Nat Resour Res 27(2):241–255
22. Zhang Q, Jiang X, Tong D, Davis SJ, Zhao H, Geng G, Ni R (2017) Transboundary health impacts of transported global air pollution and international trade. Nature 543(7647):705–709
23. Du X, Kong Q, Ge W, Zhang S, Fu L (2010) Characterization of personal exposure concentration of fine particles for adults and children exposed to high ambient concentrations in Beijing, China. J Environ Sci 22(11):1757–1764
24. Soh PW, Chang JW, Huang JW (2018) Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations. IEEE Access 6:38186–38199
25. Yi X, Zhang J, Wang Z, Li T, Zheng Y (2018) Deep distributed fusion network for air quality prediction. In Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pp 965–973
26. Zhang Y, Wang Y, Gao M, Ma Q, Zhao J, Zhang R, Huang L (2019) A predictive data feature exploration-based air quality prediction approach. IEEE Access 7:30732–30743
27. Iskandaryan D, Ramos F, Trilles S (2020) Air quality prediction in smart cities using machine learning technologies based on sensor data: a review. Appl Sci 10(7):2401
28. Xue H, Bai Y, Hu H, Xu T, Liang H (2019) A novel hybrid model based on TVIW-PSO-GSA algorithm and support vector machine for classification problems. IEEE Access 7:27789–27801
29. Mishra M (2019) Poison in the air: Declining air quality in India. Lung India Off Org Indian Chest Soc 36(2):160
30. Bishop CM (2006) Pattern recognition and machine learning. Springer, New York
31. Packtpub (2018) Machine Learning algorithms. Available online: https://www.packtpub.com/in/big-data-and-business-intelligence/machine-learning-algorithms-second-edition. Accessed on 9 Dec 2019
32. Longadge R, Dongre S (2013) Class imbalance problem in data mining review. arXiv:1305.1707
33. He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21(9):1263–1284
34. Gao M, Hong X, Chen S, Harris CJ (2011) A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems. Neurocomputing 74(17):3456–3466
35. Kubat M, Matwin S (1997) Addressing the curse of imbalanced training sets: one-sided selection. In: Icml, vol 97, pp 179–186

36. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
37. Liu XY, Wu J, Zhou ZH (2009) Exploratory undersampling for class-imbalance learning. IEEE Trans Syst Man Cybern Part B (Cybernetics) 39(2):539–550
38. Prati RC (2012) Combining feature ranking algorithms through rank aggregation. In: The 2012 international joint conference on neural networks (IJCNN), pp 1–8. IEEE
39. Gao M, Hong X, Chen S, Harris CJ (2012) Probability density function estimation based over-sampling for imbalanced two-class problems. In: The 2012 international joint conference on neural networks (IJCNN), pp 1–8, IEEE
40. Gu Q, Cai Z, Zhu L, Huang B (2008) Data mining on imbalanced data sets. In: 2008 International Conference on advanced computer theory and engineering (pp 1020–1024). IEEE
41. Sun Y, Kamel MS, Wong AK, Wang Y (2007) Cost-sensitive boosting for classification of imbalanced data. Pattern Recogn 40(12):3358–3378
42. Zhang Y, Wang D (2013) A cost-sensitive ensemble method for class-imbalanced datasets. In Abstract and applied analysis, vol 2013, Hindawi
43. Wang BX, Japkowicz N (2010) Boosting support vector machines for imbalanced data sets. Knowl Inf Syst 25(1):1–20
44. Batuwita R, Palade V (2010) FSVM-CIL: fuzzy support vector machines for class imbalance learning. IEEE Trans Fuzzy Syst 18(3):558–571
45. Cano A, Zafra A, Ventura S (2013) Weighted data gravitation classification for standard and imbalanced data. IEEE Trans Cybern 43(6):1672–1687
46. Wu G, Chang EY (2003) Class-boundary alignment for imbalanced dataset learning. In: ICML 2003 workshop on learning from imbalanced data sets II, Washington, DC, pp 49–56
47. Wu G, Chang EY (2005) KBA: Kernel boundary alignment considering imbalanced data distribution. IEEE Trans Knowl Data Eng 17(6):786–795
48. Oh S, Lee MS, Zhang BT (2010) Ensemble learning with active example selection for imbalanced biomedical data classification. IEEE/ACM Trans Comput Biol Bioinf 8(2):316–325
49. Liu Y, Yu X, Huang JX, An A (2011) Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. Inf Process Manag 47(4):617–631
50. Ertekin S, Huang J, Giles CL (2007) Active learning for class imbalance problem. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp 823–824
51. Fu J, Lee S (2013) Certainty-based active learning for sampling imbalanced datasets. Neurocomputing 119:350–358
52. Kyrkilis G, Chaloulakou A, Kassomenos PA (2007) Development of an aggregate air quality index for an urban mediterranean agglomeration: relation to potential health effects. Environ Int 33(5):670–676
53. Chelani AB, Rao CC, Phadke KM, Hasan MZ (2002) Formation of an air quality index in India. Int J Environ Stud 59(3):331–342
54. Fan S, Hazell PB, Thorat S (1999) Linkages between government spending, growth, and poverty in rural India (Vol 110). Intl Food Policy Res Inst
55. Deswal S, Verma V (2016) Annual and seasonal variations in air quality index of the national capital region, India. Int J Environ Ecol Eng 10(10):1000–1005
56. CPCB (2020) Dataset: https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing/data.
57. Maratea A, Petrosino A, Manzo M (2014) Adjusted F-measure and kernel scaling for imbalanced data learning. Inf Sci 257:331–341
58. Vapnik VN (1995) The nature of statistical learning. Theory
59. Wang L (Ed.) (2005) Support vector machines: theory and applications (Vol 177). Springer, New York
60. Foody GM, Mathur A (2004) Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification. Remote Sens Environ 93(1–2):107–117
61. Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.
62. Huang H, Xu H, Wang X, Silamu W (2015) Maximum F1-score discriminative training criterion for automatic mispronunciation detection. IEEE/ACM Trans Audio Speech Lang Process 23(4):787–797
63. Buckland M, Gey F (1994) The relationship between recall and precision. J Am Soc Inf Sci 45(1):12–19
64. Wikipedia (2021) Confusion matrix. https://en.wikipedia.org/wiki/Confusion_matrix
65. Hastie T, Rosset S, Zhu J, Zou H (2009) Multi-class adaboost. Stat Interface 2(3):349–360
66. Schapire RE (2013) Explaining adaboost. In: *Empirical inference* (pp 37–52). Springer, Berlin
67. Schapire RE, Freund Y (2013) Boosting: foundations and algorithms. Kybernetes
68. Pal SK, Mitra S (1992) Multilayer perceptron, fuzzy sets, classifiaction
69. Tang J, Deng C, Huang GB (2015) Extreme learning machine for multilayer perceptron. IEEE Trans Neural Netw Learn Syst 27(4):809–821
70. Chen MS, Manry MT (1993) Conventional modeling of the multilayer perceptron using polynomial basis functions. IEEE Trans Neural Netw 4(1):164–166
71. Bustamante C, Garrido L, Soto R (2006) Comparing fuzzy naive bayes and gaussian naive bayes for decision making in robocup 3d. In: Mexican International Conference on Artificial Intelligence, Springer, Berlin, pp 237–247
72. Griffis JC, Allendorfer JB, Szaflarski JP (2016) Voxel-based Gaussian naïve Bayes classification of ischemic stroke lesions in individual T1-weighted MRI scans. J Neurosci Methods 257:97–108
73. Wu J, Coggeshall S (2012) Foundations of predictive analytics. CRC Press
74. Ruggieri M, Plaia A (2012) An aggregate AQI: comparing different standardizations and introducing a variability index. Sci Total Environ 420:263–272
75. Friedman JM (1996) The effects of drugs on the fetus and nursing infant: a handbook for health care professionals. Johns Hopkins University Press, Baltimore
76. Cleland JG, Van Ginneken JK (1988) Maternal education and child survival in developing countries: the search for pathways of influence. Soc Sci Med 27(12):1357–1368
77. Anderson JO, Thundiyil JG, Stolbach A (2012) Clearing the air: a review of the effects of particulate matter air pollution on human health. J Med Toxicol 8(2):166–175