Automatika

Journal for Control, Measurement, Electronics, Computing and Communications

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/taut20

Optimal progressive classification study using SMOTE-SVM for stages of lung disease

R. Sujitha & B. Paramasivan

To cite this article: R. Sujitha & B. Paramasivan (2023) Optimal progressive classification study using SMOTE-SVM for stages of lung disease, Automatika, 64:4, 807-814, DOI: 10.1080/00051144.2023.2218167

To link to this article: https://doi.org/10.1080/00051144.2023.2218167

AUTOMATIKA
2023, VOL. 64, NO. 4, 807–814
https://doi.org/10.1080/00051144.2023.2218167

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Optimal progressive classification study using SMOTE-SVM for stages of lung disease

R. Sujitha[a] and B. Paramasivan[b]

[a]Department of Computer Science and Engineering, PSNA College of Engineering and Technology, Dindigul, India; [b]Department of Information Technology, National Engineering College (Autonomous), Kovilpatti, India

ABSTRACT

Data used in big data applications are typically kept in decentralized computing resources in the real world, which has an impact on the design of artificial intelligence algorithms. When there are significantly more observations from one class than from another, the dataset is said to be imbalanced. Therefore, in this work, the study elaborates the model as SMOTE-SVM which resolves imbalance issues in sampling the data and improves overall accuracy to 94%. The model deploys K-nearest neighbours to compute the difference between samples and to balance the samples, it computes the kernel space. Further, to optimize the classification, GWO optimizer merges with SMOTE-SVM to achieve enhanced performance. GWO (Grey Wolf Optimizer) induces greedy selection to perform optimization among classification. It is important to remember that grey wolves have a flexible social structure that might change the hierarchy. As the mobilization continues, the grey wolves are reconstructed with the distance between them and their prey, or more specifically, in accordance with the resultant value of the fitness. In addition, to prove the efficiency, the following performance metrics are measured-Overall Accuracy, Classification Accuracy, AUC and ROC.

ARTICLE HISTORY
Received 20 February 2023
Accepted 20 May 2023

KEYWORDS
Classification; optimization; grey-wolf optimizer; minority samples; lung cancer

1. Introduction

When a malicious cancer progresses inside the lungs, such as bronchus, it affects a small pipe that connects inside the lungs. Lung cancer, similar to other cancer types, has the ability to spread (metastasize) to other parts of the body. In this situation, cancer that starts in the lungs usually spreads to the brain, bones, adrenal glands and liver through one of three mechanisms: directly through the blood vessel, or lymphatic extension. Shortest and direct connection directs happens when a tumour grows fast in size to the point that spreads to other parts and further to other organ or structure. According to the American Cancer Society, lung cancer has become the most common type of cancer in the United States, (approximately 22 K per year). Lung cancer is not only the most frequent cancer, but it is also the most complicated one for further treatment. As a result, lung cancer is the most lethal cancer in the United States, with the statistics indicating that approximately 16 K people die every year. Lung cancer is tough to treat and cure, although it can be avoided in the majority of cases. One can nearly completely eliminate his/her risk of contracting the disease by making lifestyle changes. Stopping smoking and eating a healthy diet rich in fresh fruits and vegetables can drastically reduce your risk of lung disease.

According to the American Cancer Society, lung cancer has become the most common type of cancer in the United States, approximately 22 K per year. Lung cancer is not only the most frequent cancer, but it's also the complicated one for further treatment. As a result, lung cancer is the most lethal cancer in the United States, with the statistics that approximately 16 K lost their life per year. Lung cancer is tough to treat and cure, although it can be avoided in the majority of cases. One can nearly completely eliminate your risk of contracting the disease by making lifestyle changes. Stopping smoking and eating a healthy diet rich in fresh fruits and vegetables can drastically reduce your risk of heart disease. In our medical services, an enormous measure of higher layered information incorporates electronic clinical records, clinical notes, clinical pictures, digital actual frameworks, clinical Internet of Things, genomic information and clinical data. Handling of those information has become the most desirable and complex issue since some data set has incomplete and missing data. Datasets with such unprocessed records can be handled and accomplished positively in large amount of datasets. Some authors begin handling of

CONTACT R. Sujitha sujitha.r24@rediffmail.com Department of Computer Science and Engineering, PSNA College of Engineering and Technology, Dindigul, Tamilnadu 624 622, India

medical services information by working on its quality. They accomplished quality through data mining. In addition, medical services information are of various qualities with various sorts. Incorporation of such information has become important for additional handling. Different reconciliation methods examined so far for the target records over clinical records, Electronic Health records and observing information. Subsequently, missed data and aggregation of those data are taken care of proficiently utilizing large information and its devices. The following subsections discuss methods to be handle in electronic medical records, clinical notes, medical images, cyber-physical systems, medical Internet of Things, genomic data and clinical decision support systems. EHR are put away in HDFS or in the cloud with respect to information handled. Data should be preprocessed by filling data and converting it into a structured format. EHR handling is made out of conversion and storage in some methods. They determine information into moving, cleaning, parting, deciphering, blending, arranging and approving EHRs. Here, this unstructured data is formatted into a Structured format (e.g. patient personality, health status medication history). Further, in the proposed model, EHR is handled utilizing Map-reduce, an analytic tool.

2. Related works

Ebenuwa et al. [1] stated that some models involve the conventional features in non-small cell lung cancer as data set and use radiomics to improve performance. One of the significant approaches crosses validated Bayesian network has been deployed to improve tumoural response and local tumour before and after radiotherapy with the conventional features of the dataset. Local control prediction was enhanced by incorporating the extended Markov Blanket (eMB) technique and the wrapper-based strategy. The main objective is to classify the tumour cells before and during radiotherapy.

He et al. [2] used several optimizers and regularization methods are used to outperform the models and apply 10-fold cross-validation for performance measures. A deep learning system that classifies lung cancer images into benign or malignant nodules is proposed. The model elaborates on a multi-view knowledge-based collaborative deep model. They used chest CT data for the analysis and classification of lung cancer. To handle large datasets and classify them accurately, one of the significant factors in the proposed model should consider the unbalancing problem. Masood et al. [3] studied about handling ECG data in parallel SVM to classify different types of Arrhythmia as well as seizures. Some authors have driven the further development of parallel SVM with other applications as deployed. They elaborated multiple sub-models with parallel SVM to classify both binaries as well as multi-class classification. Some applications like multiple sub-models can classify the same datasets for both binary and multi-class with optimized parameters. Some models can classify multi-class efficiently. Some models such as AWSMOTE developed by wang et al. [4] are advanced strategies in SMOTE.

Additionally, the suggested approach enables improving the SVM classifier-based on the classification requirements. The SMOTE-SVM algorithm has been designed based on the study [5]. Mei [6] proposed a classifier for hyperspectral images with Gaussian mixture models after applying the forward feature selection method. Since hyperspectral images consist of multiple bands, band reduction achieves better feature selection than dimensionality reduction. Gao [7] Similar concepts developed in combination with PSO. Several applications are developed using SMOTE and also different algorithms have been developed to achieve balancing among classes. Concerning Naseriparsa et al. [8] they predict lung cancer by screening the conventional features. Some studies show their uniqueness in performing well in radiomics features. Both these types of features are studied well in high-dimensional data. However, some of these studies work only with limited features. Their studies are suited for conventional features. Also, accuracy and AUC score are about 75.35% and 0.75 respectively, which is slightly lower than expected measures. Nimankar [9] have several data encompassed imbalance in major and minor samples. To balance the data with major and minor samples, several algorithms are discussed here. Generally, the data imbalance problem is solved by sampling techniques. Sampling resides on oversampling (increase the minority class samples by replication) and under-sampling (reduces the majority samples and equals minority samples) [11–15].

3. Approaches and methodology

To boost the proportion of less presented examples in a data collection used for machine learning, the Synthetic

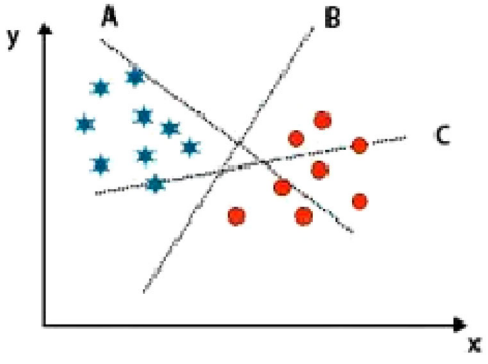


Figure 1. SVM classification.

Minority Oversampling (SMOTE) technique is used. Rather than merely duplicating current cases, this is a superior strategy for increasing the number of cases. When working with an unbalanced dataset, SMOTE is required. A dataset may be unbalanced for a variety of reasons, including the following: (i) the population may have very few members of the target group; (ii) the dataset may be challenging to gather. When the class that wants to study is underrepresented in the dataset, the study should use SMOTE, to put it simply. A balanced increase in the number of observations in the data collection can be achieved using this analytical strategy. Note that the quantity of the majority of situations is unaffected by this implementation. Since the approach considers examples of all the features for each target class and its closest neighbours, the newly generated instances are not just duplicates of the minority class that already exists. This strategy broadens the range of features that are accessible for each class and gives the examples a more generic appearance. SMOTE uses the dataset as input at the end, but it does nothing more than increase the percentage of the minority class in the data.

Table 1. Dataset accuracy of SMOTE.

Dataset	Accuracy (%)
Sputum images	93
Lung cancer	95
Thoracic	91
Overall	94
EGWO-KELM	93.42

3.1. SMOTE-SVM working principle

SMOTE can be combined with any classification techniques or single layered approach. It works better for imbalanced classification. To achieve optimal performance in both linear and non linear data, SMOTE-SVM brings a unified model. This algorithm allows reducing the time expenditures for the search of the optimum parameters values of the SMOTE algorithm.

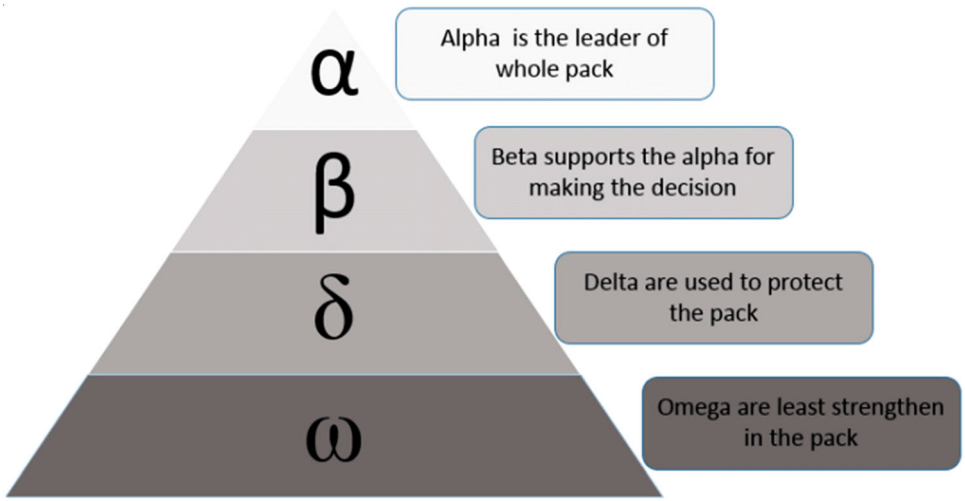


Figure 2. Hierarchy of grey wolves.

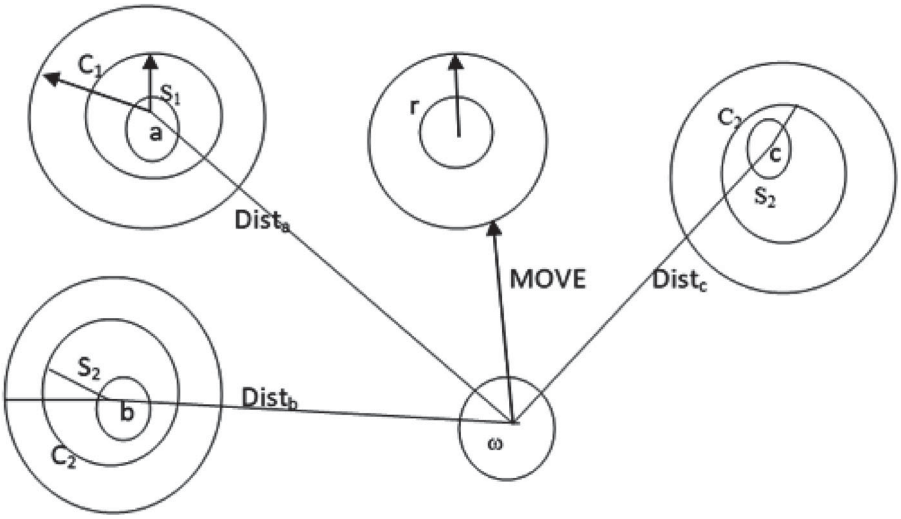


Figure 3. Operation of greedy selection.

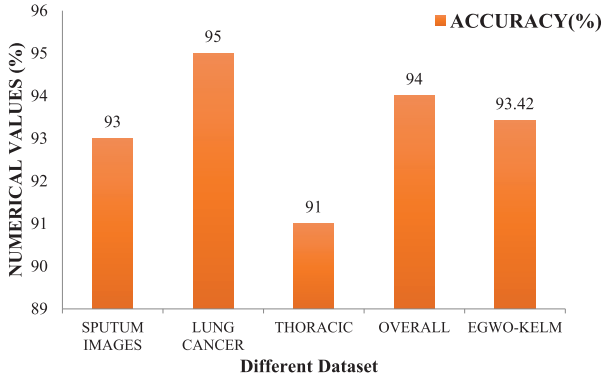


Figure 4. Graphical illustration of SMOTE accuracy.

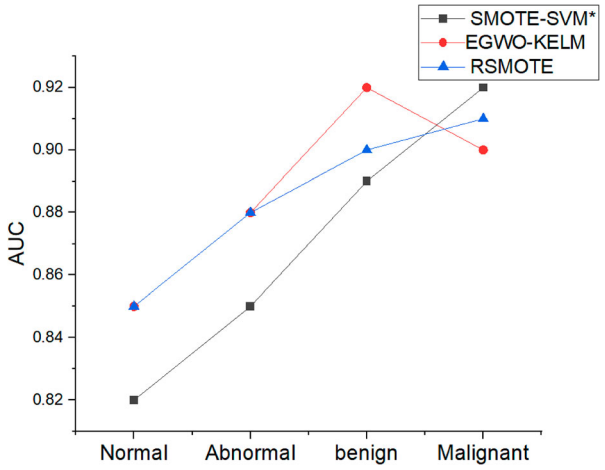


Figure 5. AUC Analysis for stages in SMOTE-SVM ∗-Proposed.

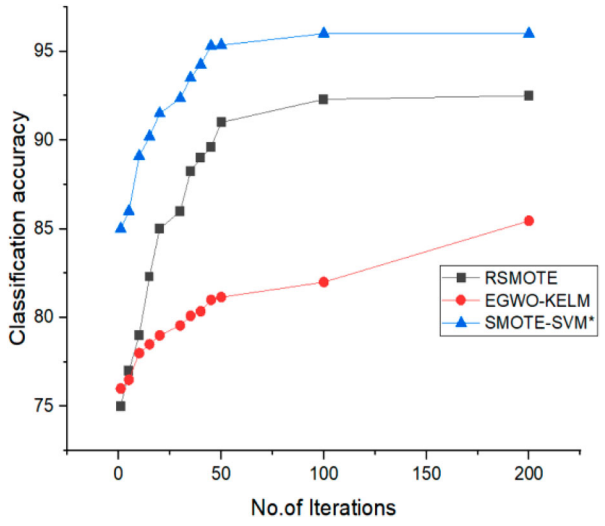


Figure 6. Classification accuracy of SMOTE-SVM ∗-Proposed.

Besides, the proposed algorithm allows increasing the data classification quality on the base of the SVM classifier as in Figure 1.

3.1.1. Sampling techniques

The application of various sampling algorithms is one of the most popular methods for addressing the issue of unbalanced datasets. The reconstruction of the classes can now be done in one of two methods. In the first scenario, a predetermined number of items belonging to the majority class are eliminated (under sampling), while in the second scenario, a predetermined range of objects belonging to the minority class are raised (over sampling). By adding (over sampling) or subtracting (under sampling) randomly chosen objects of the respective classes from the training set, random sampling is carried out. As a result, either an oversampling or an undersampling occurs. Thus, as a result of oversampling, samples of minority class and duplicate copies of the classes are made equal. Additionally, the time needed to create the classifier may rise due to object duplication. Oversampling does have the benefit of preserving all of the data, though. while deploying in the imbalance data, the study demonstrates that oversampling performs better than undersampling. Combining with random sampling, there is also a lot of usage for the unique methods that take out randomness from reconstruction of the classes.

The SMOTE sampling method in data is one of the useful specific algorithms for reconstructing the classes' balance in the situation of oversampling. The samples of minority classes are repeated in a traditional oversampling procedure. Despite the fact that the amount of data is increased, the machine learning algorithms are efficient and so the K-nearest neighbour technique is used by SMOTE to function to develop the samples. SMOTE begins by randomly selecting data from the minority class, after which the data's k-nearest neighbours are determined. Synthetic samples are generated from randomly chosen samples of k-nearest neighbour.

3.2. SMOTE-SVM algorithm

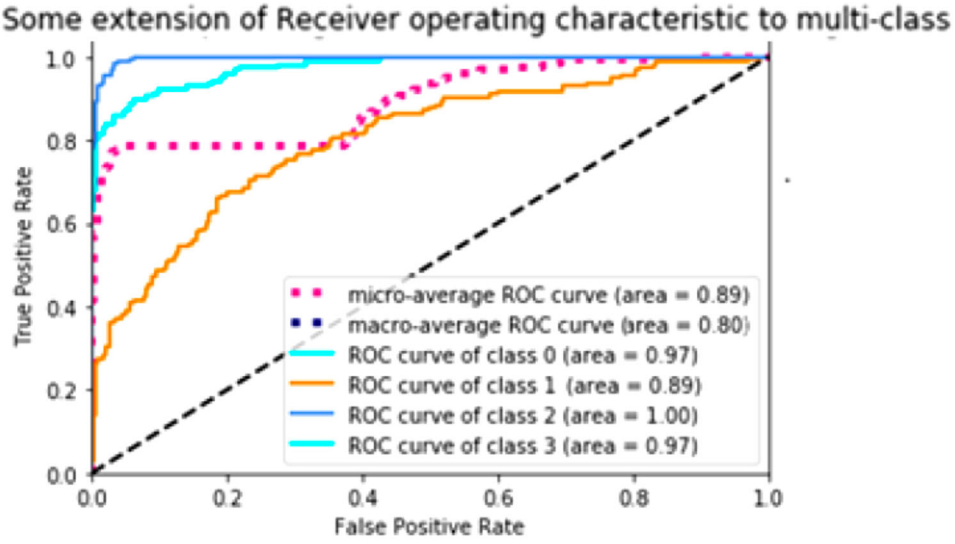
To boost the proportion of less presented examples in a data collection used for machine learning, the Synthetic Minority Oversampling (SMOTE) technique is used. Rather than merely duplicating current cases, this is a superior strategy for increasing the number of cases. When working with an unbalanced dataset, SMOTE is required. A dataset may be unbalanced for a variety of reasons, including the following: (i) the population may have very few members of the target group; (ii) the dataset may be challenging to gather.

As said above, the slack variable significance shows in Equation (1)

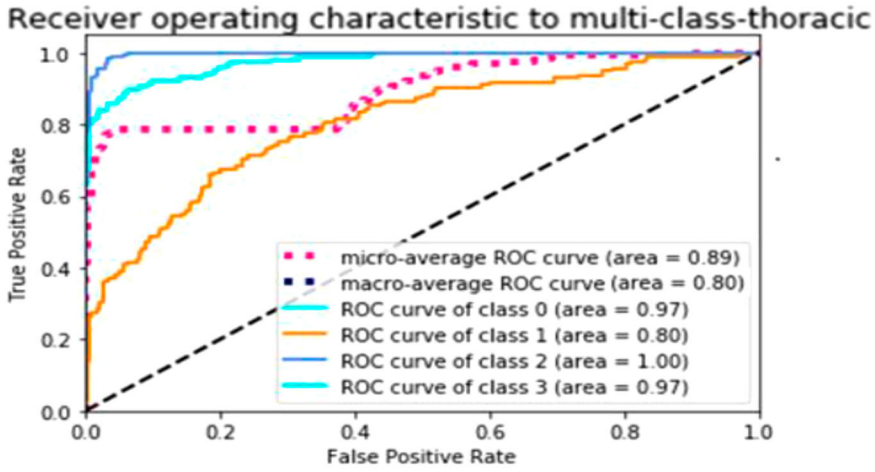
$$\min_{w,b} W.W + C\sum_{j}\xi_j(W.X_j + b)y_j \geq 1 - \xi_j, \forall_j \xi \geq 0 \tag{1}$$

where the condition for Slack penalty, $C > 0$;

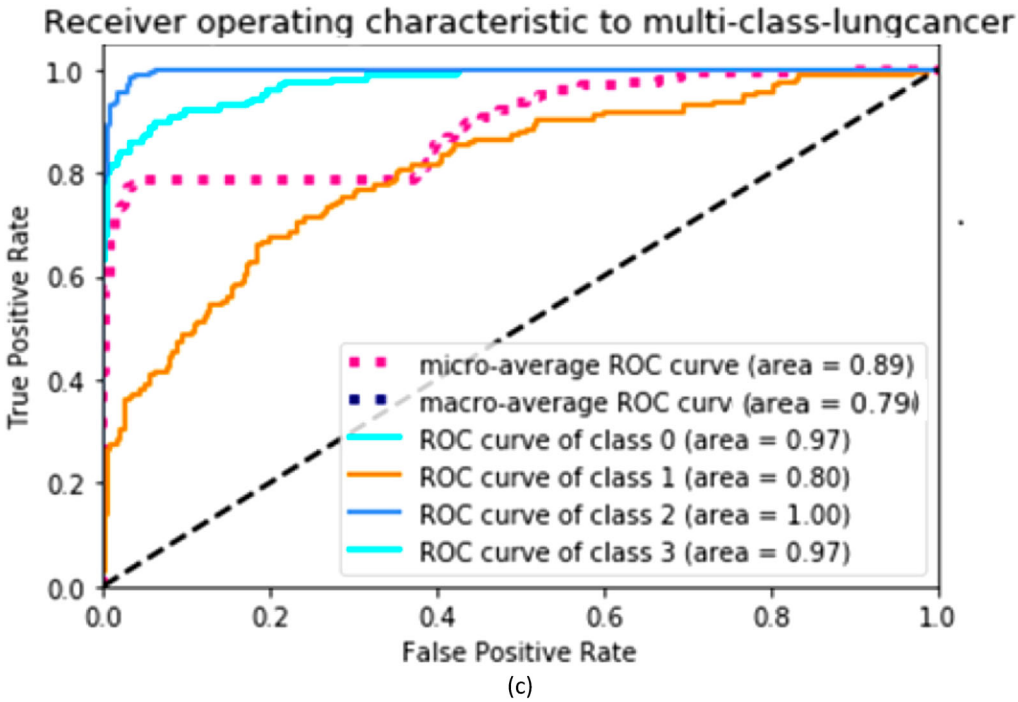
SMOTE is a scalable method since it outperforms even as a single approach and proves its uniqueness when combining other algorithms. SMOTE-SVM introduces a unified model to get the best performance in both linear and nonlinear data. With the use of this approach, it can reduce the amount of time



(a)



(b)



(c)

Figure 7. (a) ROC for sputum images dataset, (b) ROC for thoracic dataset and (c) ROC for lung cancer dataset.

SMOTE_SVM Algorithm

Step 1: Randomly sample dataset from minority samples
Step 2: Compute D_i (feature vector) and DK_{nn} (k-nearest neighbour from the minority samples)
Step 3: compute the difference between feature vectors obtained and minority samples generated using K-nn.
Step 4: product the values of output by "r" (a random number between 0 and 1)
Step 5: Add the output to the feature vector D_i to select a new point $D_1 \ldots D_i$ on the line segment between feature vectors
Step 6: Repeat steps from 1 to 4 to identify new feature vectors
Step 7: Compute kernel space using SVM and optimizer GWO
Step 8: Train the SVM with the balanced new dataset and apply optimization for better classification

needed to find the SMOTE algorithm's ideal parameter values. Additionally, the suggested approach enables improving the SVM classifier-based on the classification requirements. The SMOTE-SVM algorithm has been designed based on the study [5].

A SMOTE based SVM classifier is used in SMOTE-SVM, a variation of the SMOTE method, to collect the samples that will be utilized to create random samples. After employing support vectors to roughly estimate the margin, the SVM algorithm is applied to the initial training dataset. As it generates synthetic data with reduced overlapping, the model perforates the data separation. All support vectors of the minority samples are connected by random lines that correspond to its neighbours. SMOTE was used in this investigation to roughly balance our dataset by equalizing the major and minor class samples.

$$DS_{syn} = DS_i + (DS_{knn} - DS_i) \times rand \qquad (2)$$

where DS_{syn} represents Synthetic Data set, DS_i represents minority samples, DS_{knn} represents k-nearest neighbours from the minority class samples and rand is the random number that varies from 0 to 1.

3.3. Optimization using GWO

Grey Wolf Optimizer (GWO) optimizes SVM classification by reclassifying with search based best values. Wang, et al. [10] viewed that, basically, model divides the wolves into four categories, named high, medium, low and none. The tiers nearest to the target, and the three top levels of the hierarchy are shown in Figure 2. The other category none ω which is the least level of hierarchy, makes up the other grey wolf individuals. The top levels in the hierarchy makes way for ω for inspire hunting. It is important to remember that grey wolves have a flexible social structure that might change the hierarchy. As the mobilization continues, the grey wolves are reconstructed with the distance between them and their prey, or more specifically, in accordance with the resultant value of the fitness, that is computed as in the algorithm.

Figure 3 depicts the operation of greedy selection.

4. Results and discussion

Since the proposed model specifically consciousness on the multi-class, the proposed model used an apache spark structure embedded with the system strategies. The following metrics are analysed for the performance.

- Overall Accuracy
- Classification Accuracy
- AUC
- ROC

4.1. Overall accuracy

SMOTE-SVM model accuracy has been compared with its related arts and proves its efficiency (Table 1). However, the dataset with high dimensional categorical data are preprocessed and optimized with GWO to improve classification accuracy as in Figure 4.

4.2. AUC analysis

Graphical representation for AUC analysis of all the stages in the proposed Model SMOTE-SVM has been compared with other models Wang [4] and Naseriparsa [8] are illustrated in Figure 5.

From the Figure 5, it is clear that EGWO-KELM has increased AUC for the stage Malignant than others. But the proposed model is improved in earlier stages while comparing with others. Overall AUC value of the proposed model is 0.92. The main objective of this study is to enhance the performance of classification by reducing misclassification rate of previous study.

4.3. Classification accuracy

In order to achieve the improvement and to override the misclassification rate of previous study, the classification accuracy has been measured in this study as in Figure 6.

AUC and classification accuracy is improved than previous study as in Figure 6. AUC of T-BMSVM is 0.89 and the improved classification SMOTE-SVM is 0.92 which achieves the enhancement in the performance as in the above figure.

4.4. ROC for multiclass

In these results, ROC for multiclass shows improvement from previous model T-BMSVM. The proposed SMOTE-SVM demonstrates its hugeness with decreased emphasis and encoding technique. It is additionally worth in estimating the showing and its qualities in the following Figure 7 in multi-class for each stage 0 to 4 in exhibiting the design and their comparing with micro and macro averages.

GWO Algorithm

Step 1: Initialize the parameters (population size, iteration)
Step 2: Find the best (X_a, X_b, X_c) positions based on the equations (deep belief network)
Step 3: For iteration $t = 1$, compute the following

$$\begin{cases} \overrightarrow{X_1} = \overrightarrow{X_a} - \overrightarrow{S_1}.\overrightarrow{\text{Dist}_a} \\ \overrightarrow{X_2} = \overrightarrow{X_b} - \overrightarrow{S_2}.\overrightarrow{\text{Dist}_b} \\ \overrightarrow{X_3} = \overrightarrow{X_c} - \overrightarrow{S_3}.\overrightarrow{\text{Dist}_c} \end{cases} \quad (1)$$

$$\begin{aligned} \overrightarrow{S_1} &= 2c.\overrightarrow{\text{rand}_4} - c \\ \overrightarrow{S_2} &= 2c.\overrightarrow{\text{rand}_5} - c \\ \overrightarrow{S_3} &= 2c.\overrightarrow{\text{rand}_6} - c \end{aligned} \quad (2)$$

$$\text{Convergence factor} c = 2 - 2\left(\frac{n^{\text{th}}\text{iteration}}{\text{Total iteration}}\right) \quad (3)$$

Step 4: Perform greedy selection to update the position and optimal solution. In order to determine the finite solution,individuals must enlarge its search area according to the global search method. The distance between ω and a, b and c is expressed by the following formula. C1, C2, C3 represents vector from a to ω, b to ω and c to ω respectively,

Dist_a = diff.bet position of a^{th} and ω with respect to C_1 (4)
Dist_b = diff.between position of b^{th} and ω with respect to C_2 (5)
Dist_c = diff.between position of c^{th} and ω with respect to C_3 (6)

Step 5: Therefore, the grey wolves carrying out the outer encirclement order approach the prey from far to near. In order to avoid the algorithm falling into the local optimum, the grey wolves performing the outer encirclement must obey the commands of α, β and δ. When ω receives the command to kill, it moves closer to the prey to update its position.

Table 2. Comparison of all metrics with several SMOTE models.

Algorithms	*F*-measure	ACC	AUC	Precision
BLSMOTE	0.92	0.88	0.9	0.951
AWSMOTE	0.91	0.92	0.925	0.95
SMOTE∗	0.89	0.94	0.92	0.95
WK-SMOTE	0.85	0.8	0.86	0.82
RSMOTE	0.86	0.9	0.91	0.89

∗-SMOTE.

ROC Values for all three datasets have been measured and it shows that all the datasets improves its performance to about 0.92 in multi-class.

AUC values and other performance metrics have been compared with all types of SMOTE such as AWSMOTE developed by Wang [4] and improved for all classes in all dataset as in Table 2.

5. Conclusion

Combining with Python and spark in LIBSVM, **t**his work launches a combined methodology to categorize the tiers of nodules. s the classification methods in SVM although simpler, it is mandatory to improve performance accuracy. The model extends the works to optimum classification using the imbalance classification method SMOTE. The proposed method oversamples the classes and balances the samples for every class. The model modifies the parameters of SMOTE to work well in multi-class classification. The purpose of SMOTE is to equalize the number of samples in every class. Then it updates the parameter for multiclass by manually labelling the strategy. The optimization algorithm combined with SMOTE removes the over-generalization problem. Phase III extends the classification with the optimization model Grey Wolf Optimization (GWO) is deployed in combination with SMOTE for SVM classification. Grey Wolf Optimization works and inherits optimal outcomes by using the model called group hunting of the grey wolves. It simulates the hunting behaviour of the grey wolf and finds n number of best solutions to optimize the margin. The fitness function of GWO determines the best solution or parameters for SVM classification. Then the prediction or classification task is performed using the optimal parameters. The best parameters are K for cross-validation, number of search agents, and number of iterations, and functions to apply to the search agent. Though it is furnished, individual dataset classification tends to increase execution time and processing time. Hence our future work will focus on integrated classification with increased set of datasets.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Ebenuwa SH, Sharif MS, Alazab M, et al. Variance ranking attributes selection techniques for binary classification problem in imbalance data. IEEE Access. 2019;7:24649–24666.
[2] He T, Jixiang G, Nan C, et al. Medi MLP: using grad-CAM to extract crucial variables for lung cancer postoperative complication prediction. IEEE J Biomed Health Inform. 2020;24(6):1762–1771.
[3] Masood A, Bin S, Po Y, et al. Automated decision support system for lung cancer detection and classification via enhanced RFCN with multilayer fusion RPN. IEEE Trans Ind Inf. 2020;16(12):7791–7801.
[4] Wang J-B, Zou C, Fu G. AWSMOTE: An SVM-based adaptive weighted SMOTE for class-imbalance learning. Sci Program. 2021: 1–18. DOI: 10.1155/2021/9947621.

[5] Mathew J, Pang CK, Luo M, et al. Classification of imbalanced data by oversampling in kernel space of support vector machines. IEEE Trans Neural Netw Learn Syst. 2018;29(9):4065–4076.
[6] Mei, X. Predicting five-year overall survival in patients with non-small cell lung cancer by relief algorithm and random forests. In 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE; 2017. p. 2527–2530; doi:10.1109/IAEAC.2017.8054479.
[7] Gao M, Hong X, Chen S, et al. A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems. Neuro Comput. 2011;74(17):3456–3466.
[8] Naseriparsa M, Al-Shammari A, Sheng M, et al. RSMOTE: improving classification performance over imbalanced medical datasets. Health Inf Sci Syst. 2020;8(22). doi:10.1007/s13755-020-00112-w.
[9] Nimankar SS, Vora D. Designing a model to handle imbalance data classification using SMOTE and optimized classifier. In: Sharma N, Chakrabarti A, Balas V, etal, editors. Data management, analytics and innovation: Proceedings of ICDMAI 2020. Vol. 1. Singapore: Springer; 2021. p. 323–334.
[10] Wang Z, Zeng Y, Liu Y, et al. Deep belief network integrating improved kernel-based extreme learning machine for network intrusion detection. IEEE Access. 2021;9:16062–16091.
[11] Maulik U. Fuzzy preference-based feature selection and semi-supervised SVM for cancer classification. IEEE Trans NANO Biosci. 2014;13(2):152–160.
[12] Ozdemir O, Russell RL, Berlin AA. A 3D probabilistic deep learning system for detection and diagnosis of lung cancer using low-dose CT scans. IEEE Trans Med Imaging. 2020;39(5):1419–1429.
[13] Rafi TH, Shubair RM, Farhan F, et al. Recent advances in computer-aided medical diagnosis using machine learning algorithms with optimization techniques. IEEE Access. 2021;9:137847–137868.
[14] Su Z, Yuntao W, Tom H, et al. Secure and efficient federated learning for smart grid with edge-cloud collaboration. IEEE Trans Ind Inf. 2022;18(2):1333–1344.
[15] Yazdani H, Cheng LL, Christiani DC. Bounded fuzzy possibilistic method reveals information about lung cancer through analysis of metabolomics. IEEE/ACM Trans Comput Biol Bioinf. 2020;17(2):526–535.