Original Research Paper

Combining SMOTE and OVA with Deep Learning and Ensemble Classifiers for Multiclass Imbalanced

Rattanawadee Panthong

School of Information and Communication Technology, University of Phayao, Thailand

Abstract: The classification of real-world problems always consists of imbalanced and multiclass datasets. A dataset having unbalanced and multiple classes will have an impact on the pattern of the classification model and the classification accuracy, which will be decreased. Hence, oversampling method keeps the class of dataset balanced and avoids the overfitting problem. The purposes of the study were to handle multiclass imbalanced datasets and to improve the effectiveness of the classification model. This study proposed a hybrid method by combining the Synthetic Minority Oversampling Technique (SMOTE) and One-Versus-All (OVA) with deep learning and ensemble classifiers; stacking and random forest algorithms for multiclass imbalanced data handling. Datasets consisting of different numbers of classes and imbalances are gained from the UCI Machine Learning Repository. The research outputs illustrated that the presented method acquired the best accuracy value at 98.51% when the deep learning classifier was used to evaluate model classification performance in the new-thyroid dataset. The proposed method using the stacking algorithm received a higher accuracy rate than other methods in the car, pageblocks, and *Ecoli* datasets. In addition, the outputs gained the highest performance of classification at 98.47% in the dermatology dataset where the random forest is used as a classifier.

Keywords: SMOTE, One-Versus-All, Multiclass Imbalanced, Deep Learning, Ensemble Classifiers

Introduction

Machine learning is a subfield of Artificial Intelligence (AI) and predictive analytics. It is developed to make computers learn by creating algorithms that can gain input data and apply statistics to forecast the output meanwhile evolving and adapting it as new data is available. Most practical machine learning techniques need supervised learning. The supervised learning method iterative predicts training data based on the class label. Classification is a supervised learning technique using machine learning algorithms and labeled training data to obtain how to determine class labels to sample from the domain. The problem of class imbalance is the main problem in machine learning such as classification problems with imbalanced datasets. Class imbalance is seen in real-world problems, including spam filtering, disease screening, fraud detection, and advertising click-throughs.

The classification in imbalanced datasets that the number of instances in the training dataset for each class label is not balanced. The unbalanced data as a training set where the majority class is greater than the minority class, for example, the medical prognosis work of detecting diseases in which the majority group of the patients is healthy or negative and detecting the diseases as unhealthy or positive cases in a minority group (Johnson and Khoshgoftaar, 2019; Hasib *et al.*, 2020). The data level approaches, the algorithm level techniques, and the hybrid methods are applied to manage class imbalance in machine learning (Leevy *et al.*, 2018). Firstly, the data level approaches are a method of preprocessing data before classifying using traditional classifiers. The main handle of multiclass imbalanced datasets of the data level techniques is to reduce the impact of skewed class distribution and to rebalance label class. Data-level approaches are suitable for the solutions to the problems of class imbalance. The algorithm-level

methods can not only illustrate particular algorithms but also adapt the procedure of the classifier learning itself. Finally, the hybrid-level techniques are combined learning algorithms. An example of a hybrid level is the boosting technique. (Mahani and Ali, 2019; Upadhyay *et al*., 2021)

In classification problems, the approaches for handling imbalanced problems are to increase the samples of a dataset. One of them is the most commonly known as Synthetic Minority Oversampling Technique (SMOTE) (Yu and Zhou, 2021). SMOTE is a very popular oversampling or up sampling technique. It is used in data preprocessing methods for handling imbalanced datasets. This method is an up-sampling method where the synthetic examples are produced for the minority class. SMOTE is applied to raise the sample of the minority class to equal the class distribution (Maheshwari *et al*., 2018; Hasib *et al*., 2020). Furthermore, it helps avoid the overfitting problem, and also it is applied to improve the effectiveness of a classification model in imbalanced datasets (Sun *et al*., 2015). Hence, the up sampling approach can be applied to solve the imbalance problem in the machine learning method (Charte *et al*., 2013).

In the machine learning fields, the multiclass imbalanced problem is one of the most challenging. There are plenty of classification problems that occur and involve multiclass type and imbalanced data. Also, real-world problems often have multiple classes: Bioinformatics, image, handwriting recognition, face recognition, speech recognition, and text. The classification problems are supervised learning methods in which each example is correlated with a class target or label. It is trained on training data that has been labeled. Multiclass problem is the classification tasks with more than two classes in the machine learning method. There are very popular two techniques, including One-Versus-All (OVA) and One-Versus-One (OVO). The approaches of multiclass classification are to change the multiple class problems into several binary problems using a heuristic method (Bolon-Canedo *et al*., 2011; Özdemir *et al*., 2021). In this study, the OVA technique is applied to split the classifying k class problems into k binary classification problems, in which each problem discriminates a given class from the other k−1 classes (Mehra and Gupta, 2013). The step of learning the classifiers divides the single class as a positive and the rest as negatives. Therefore, the present work combined methods SMOTE and OVA and has the as main objective to handle multiclass imbalanced datasets, and improve the classification model.

In summary, this study presented the techniques of handling imbalanced datasets for multiclass classification, including SMOTE approach and OVA strategy based on deep learning and ensemble classifiers; stacking, and random forest algorithms.

These methods are applied to improve the model of prediction and to increase accuracy. Moreover, they are taken to handle multiclass imbalanced datasets.

Background and Related work

This section reviews the oversampling approach and OVA strategy for multiclass imbalanced classification problems. There are several novel oversampling methods for imbalanced datasets. An imbalanced classification problem is the issue of classification when the training dataset has an unbalanced distribution of classes. A real-world imbalanced classification problem may have more two class labels or multiclass classification.

SMOTE Approach

Many researchers have proposed using oversampling methods for handling multiple class and unbalanced datasets. The sampling approach is a preprocessing of training sets that are used to create balanced a dataset and adjust the prior distribution for minority and majority classes. There are two approaches including under-sampling and oversampling methods. SMOTE strategy generates synthetic instances to balance class distribution by random replicating in-stances of minority classes (Abd Elrahman and Abraham, 2013; Maheshwari *et al*., 2018; Hasib *et al*., 2020; Almayyan, 2021; Yu and Zhou, 2021).

Majumder *et al*. (2020) proposed a method of handling multiclass imbalanced problems using geometry-based information sampling and class prioritized synthetic data creation (GICaPS). The combination of oversampling and undersampling approaches is applied to improve the class division and increase the diversity of samples within each class. The oversampling technique is preprocessing class unbalanced training set where it adds more numbers of examples in the minority class by the randomized method. Furthermore, it tends to balance the number of instances for the majority and minority classes in the training set. This method is a very effective technique for the classification of imbalanced datasets.

The up-sampling approaches for handling multiclass imbalanced datasets were presented by Sáez *et al*. (2016) The purpose was to analyze subsets of specific instances in different classes. In this study, the training sets 21 multiclass datasets taken from the UCI repository. The results showed the highest average accuracy (72.56%) in all datasets when C4.5 is used as a classifier.

Özdemir *et al*. (2021) proposed a method of imbalanced hyperspectral images for the classification using SMOTE that relied on deep learning algorithms. In this experimentation, the IEEE Dataport was used as a training dataset. The results obtained the highest accuracy rate (96.49%, 95.64%, 93.38%) in the multiclass hyperspectral image dataset when smote balanced with 5-fold cross-validation was applied.

Waqar *et al.* (2021) proposed the method of SMOTE technique depended on a deep learning algorithm for the prediction of a heart attack. SMOTE is used to handle imbalanced datasets. This study used datasets from UCI. The experimental results illustrated that SMOTE based on an artificial neural network algorithm when tuned properly outperformed all other models. Also, this method obtained good performance when it is used to classify heart failure. The results gained the best average accuracy (96%) when using a SMOTE-based artificial neural network.

In addition, Yuan *et al.* (2018) developed the ensemble method of the deep learning algorithm. It is used to handle imbalanced problems. The stratified under-sampling is used to balance all classes. The results illustrate the best performance of the method for imbalanced multiclass classification problems. Furthermore, the highest accuracy improvement is at 24.7% and the proposed approaches decreased the computational cost.

In summary, the oversampling approach is preferred over the undersampling techniques because the undersampling approach tends to remove instances from data that may carry some important information.

One-Versus-All (OVA) Strategy for Multiclass Classification

OVA is also called One-Against-All (OAA), a popular heuristic approach. It separates the multiclass classification problem into several binary classification ones. The binary classifier is trained on each binary classification problem and then the prediction is created using the model, with the highest confidence score (Abd Elrahman and Abraham, 2013). The classifier predicts instance x as class label y, providing a maximum probability score; the calculation of the final decision function was measured using Eq. (1) (Ghanem *et al.*, 2010):

$$discri\min ant\ function: F(x) = \mathop{\arg\max}_{i=1,\dots,k} f_i \qquad (1)$$

Where assigns the test instance to the class with the highest output value.

Other researchers have proposed a combination of SMOTE and OVA strategies for multiclass imbalanced datasets. Puttiporn and Yaowares proposed an approach of adaptive synthetic sampling technique (ADASYN) and SMOTE approach for oversampling in the dataset. The one-vs-one and one-vs-all with the Gentleboost algorithm are applied to handle imbalanced data with multiple classes. The assessment of the elderly's knee osteoarthritis incidence dataset was obtained from Ban Han Sub-District Health Promoting Hospital, Thasala district, Nakhon Si Thammarat province. The experimental results indicated that the ADASYN method and OVO strategy received the highest accuracy rate of 97.31% (Puttiporn and Yaowares, 2019).

The SMOTE approach for imbalanced big data using Random Forest was presented by Bhagat and Patil (2015). In this study, the proposed method for classification of multiclass imbalanced data. This approach composes of two stages: In the first stage, OVA and OVO strategies are used for splitting training datasets into subsets of binary classes. In the second stage, the SMOTE technique is used to balance the training set. The Random Forest (RF) algorithm is used to classify the predictive model. The SMOTE method is adjusted to huge data using MapReduce and to handle a huge dataset. The different datasets were gained from the UCI repository for the experiment. Furthermore, this method is implemented on Apache Hadoop and Apache Spark platforms. The experimental results acquired illustrated that the proposed method had better performance than other methods.

The differences between the proposed method and the previous work (Puttiporn and Yaowares, 2019) are as the following. Firstly, in preprocessing step, the present work combined SMOTE and OVA strategies based on deep learning and ensemble methods for multiclass imbalanced datasets. In this step, three different learning algorithms (deep learning, stacking, and random forest) were used to evaluate subsets of each class, whereas previous work [xx] proposed the method of SMOTE based on OVO and OVA based on the Gentleboost algorithm. Lastly, the classification and the assessment used deep learning and ensemble methods via 10-fold cross-validation. These methods can help improve good generalization effectiveness of classification.

Performance Measures for Multiclass Imbalanced Classification

In this study, the accuracy, kappa, sensitivity, and specificity are used to assess the model of classification. The performance of multiclass imbalanced classification is typically evaluated with a confusion matrix as illustrated in Table 1.

Accuracy

The accuracy value is used to measure the classification evaluation on the datasets and to compare the accuracy of the classification models. It is the percentage of correctly classifies instances out of all instances. The efficiency of the classifier is assessed in terms of the following measures: True Positive (TP) is the number of positive samples that are precisely classified. True Negative (TN) is the number of negative samples that are precisely classified. False Negative (FN) is the number of positive samples that are incorrectly classified as negative. False Positive (FP) is the number of negative samples that are incorrectly classified as positive (Farid *et al.*, 2014):

$$Accuracy = \frac{(TP + TN)}{(TP + FN + FP + TN)} \qquad (2)$$

Sensitivity

Sensitivity is also called the recall or the true positive rate. It is the proportion of true positives, correctly classified as positive class labels. It is applied to improve model classification for multiclass problems. Sensitivity based on the micro average method is applied to measure the effectiveness of the classification where the training set varies in size or imbalanced problem (Panthong and Srivihok, 2019). The sensitivity value can be calculated as:

$$Sensitivity = \frac{\sum_{i=1}^{k} tp_i}{\sum_{i=1}^{k} tp_i + fn_i} \qquad (3)$$

Specificity

Specificity is the ratio of true negatives, correctly classified as negative class labels. Specificity based on the micro average method is applied to measure the effectiveness of the classification model where the dataset varies in size or imbalanced problem (Panthong and Srivihok, 2019). The specificity can be computed as shown below:

$$Specificity = \frac{\sum_{i=1}^{k} tn_i}{\sum_{i=1}^{k} (tn_i + fp_i)} \qquad (4)$$

Cohen's Kappa

Cohen's Kappa or Kappa is like classification accuracy, except that it is depended on the confusion matrix. The Kappa is a more useful measure to use on problems having an imbalance in the classes. When working with an imbalanced dataset, this value is more informative than overall accuracy. Cohen's kappa coefficient is a measurement method that can handle both multiclass and imbalanced class problems. Cohen's kappa is calculated with the following formula (De Raadt *et al.*, 2019):

$$K = \frac{p_0 - p_e}{1 - p_e} \qquad (5)$$

Where, p_0 is the overall accuracy of the predictive model and p_e is a measurement of the consent between the model predictions and the actual class values.

Materials and Methods

The SMOTE approach and OVA strategy are used to handle multiclass imbalanced datasets. SMOTE method is used to create instances for the minority class. OVA with different classifiers were applied to improve the accuracy of model classification. Combining SMOTE approach and

OVA strategy with different classifiers used in the present study and associated abbreviations. Moreover, these methods have a mechanism to avoid overfitting, since typically cross-validation measures of predictive accuracy are used. The framework for SMOTE approach and OVA strategy with deep learning and ensemble classifiers; stacking and random forest algorithms for multiclass imbalanced dataset handling is illustrated in Fig. 1.

The researcher used the software RapidMiner Studio 9 (https://rapidminer.com/platform/educational/) on windows 10 (64-bit) via desktop computer or notebook. Furthermore, the experiments were performed under the following computer specification:

1. Computer Intel Core i5-6300U CPU 2.40 GHz
2. RAM 8 GB
3. HDD 500 GB

The Process of Combining Smote Approach and Ova Strategy with Deep Learning and Ensemble Classifiers for Multiclass Imbalanced Data Handling

Initial 9 training datasets from UCI datasets.
The method preprocessing approach contains three steps:

Step1: The up-sampling approach with SMOTE is applied to balance a training dataset, that its synthesis instances for the minority class. In this study, the oversampling approach uses the SMOTE up sampling operator. In this step, the training set is filtered to only consider instances of the minority class

Step 2: The dataset is modified for the OVA strategy. The OVA technique is applied to divide each class of training sets where one class is set as positive, the rest classes are assigned negative

Step 3: Select the final model with three different classifiers. In this study work, deep learning and ensemble classifiers (stacking, and random forest algorithms) are used to evaluate the classification model

The deep learning algorithm is a subcategory of machine learning methods. It is supervised training for regression and classification tasks. The deep learning method is applied to multiclass classification problems. In addition, it can help evaluate the performance of classification in supervised learning (Candel and LeDell, 2022). In this study, the learning algorithm with H2O is based on a multilayer feedforward Artificial Neural Network (ANN) for predictive modeling that is trained using Backpropagation (BP). The ANN model or Multilayer Perceptron (MLP) is the most widely applied deep neural network. Deep learning model to identify different classes of the training dataset.

A stacking learning algorithm or stacked generalization is a machine learning ensemble where the models are combined using another machine

learning algorithm. The main concept is to train learning algorithms with a dataset and then produce a new training set with these models. Then, this new training set is used as input data for the combiner of the learning algorithm of predictive models. stacked generalization is trained on a separate dataset to deduce the biases of the learning set and avoid overfitting. The base classifier often consists of different learning algorithms. Therefore, stacking ensembles are often heterogeneous (Xie *et al.*, 2022).

A random forest algorithm is an ensemble of a certain number of random trees. This approach is an extension of the bagging technique based on the predictions of the decision trees. It is a very popular decision tree ensemble. The random forest technique is a type of supervised training. It is used widely in classification problems. This method has proven its effectiveness on a wide range of different predictive modeling problems. The random forest algorithm makes a decision tree prediction more efficient and increases the accuracy and the robustness of the classification model (Bisht *et al.*, 2016; Putri *et al.*, 2021).
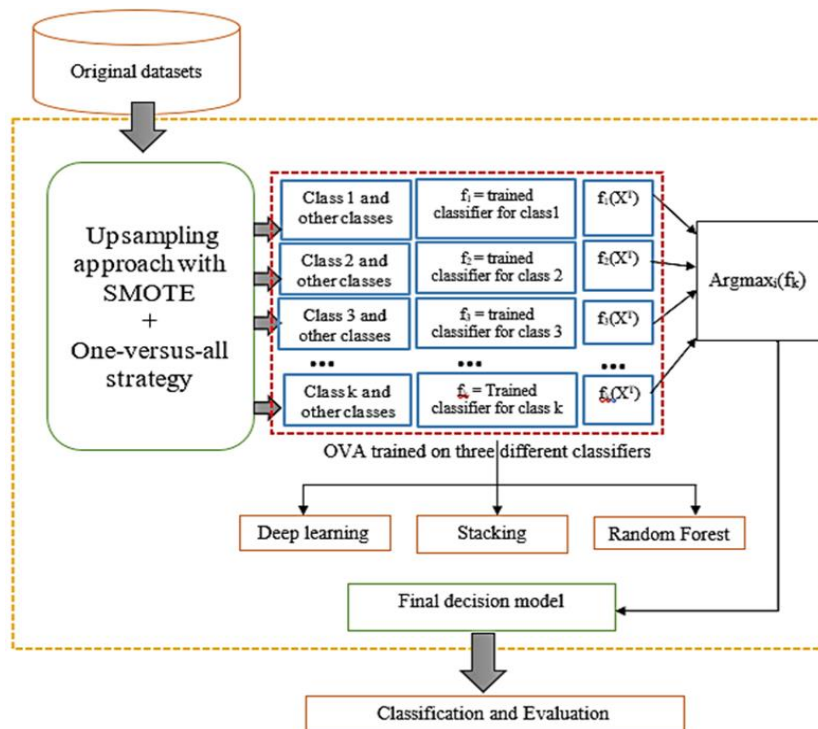

Fig. 1: Framework of combining SMOTE approach and OVA strategy with different classifiers for multiclass imbalanced data handling

RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	Type
1.52101	13.64	4.49	1.1	71.78	0.06	8.75	0	0	1
1.51761	13.89	3.6	1.36	72.73	0.48	7.83	0	0	1
1.51618	13.53	3.55	1.54	72.99	0.39	7.78	0	0	1
1.51766	13.21	3.69	1.29	72.61	0.57	8.22	0	0	1
1.51742	13.27	3.62	1.24	73.08	0.55	8.07	0	0	1
1.51596	12.79	3.61	1.62	72.97	0.64	8.07	0	0.26	1
1.51743	13.3	3.6	1.14	73.09	0.58	8.17	0	0	1
1.51756	13.15	3.61	1.05	73.24	0.57	8.24	0	0	1
1.51918	14.04	3.58	1.37	72.08	0.56	8.3	0	0	1
1.51755	13	3.6	1.36	72.99	0.57	8.4	0	0.11	1
1.51571	12.72	3.46	1.56	73.2	0.67	8.09	0	0.24	1

Fig. 2: An example of a glass dataset (Source: https://www.kaggle.com/datasets/uciml/glass)

Table 1: Confusion matrix for multiclass classification

		Predicted Class			
		Class1	Class2	…	Class k
True Class/ Actual Class	Class1	C11	C12	…	C1k
	Class2	C21	C22	…	C2k
	…	…	…	…	…
	Class k	Ck1	Ck2	…	Ckk

Table 2: The dataset description used in the experiment

Dataset	No. of instances	No. of classes	No. of instances per class	Imbalance ratio (IR)
Yeast	1484	10	244/429/ 463/44/35/51/163/30/ 20/5	92.60
Glass	214	7	70/76/17/13/9/29	8.44
Car	1728	4	1210/384/65/69	18.62
Dermatology	366	6	112/61/72/49/52/20	5.60
New-Thyroid	215	3	150/35/30	5.00
Pageblocks	5473	5	4913/329/28/88/115	175.46
Vertabral3c	310	3	150/100/60	2.50
Cleveland	297	5	160/54/35/35/13	12.31
Ecoli	335	8	142/77/52/35/20/5/2/2	71.00

Classification and Evaluation

In this study, the three classifiers (deep learning, stacking, and random forest) were used for the classification and evaluation models. Furthermore, in this step 10-fold cross-validation is applied. Cross-validation estimates the statistical performance of a learning algorithm by dividing data into separate training and test sets. Moreover, it increases the accuracy rate of datasets. This study uses RapidMiner studio version nine enterprise educational edition for model training and testing. It is used as a tool for preprocessing methods, training classifiers, and performance evaluation. In this study, the proposed method is compared to common SMOTE and without SMOTE. Three learning algorithms; deep learning, stacking, and the random forest are used for the classification and evaluation models. All datasets are tested through 10-fold cross-validation.

Datasets

The datasets for this experiment using 9 multiclass imbalanced training sets were received from the UCI Machine Learning Repository and all datasets were downloaded on the website (http://archive.ics.uci.edu/ml). The data are described in Table 2. Furthermore, the method of the SMOTE approach and OVA strategy with three different classifiers used in the current research and correlated abbreviations showed in Table 3.

Here is a dataset example and an access link as illustrated in Fig. 2.

Results and Discussion

The experimental results of combining SMOTE approach and OVA strategy with three different classifiers for multiclass imbalanced dataset handling were presented.

The three different classifiers, namely deep learning, stacking, and random forest algorithm with 10-fold cross-validation were used for evaluating the performance of the prediction. The proposed method was compared with non-preprocessing and preprocessing with SMOTE. Furthermore, the accuracy, sensitivity, and specificity of the training sets of the 9 benchmark datasets were used to evaluate the performance classification.

Tables 4, 5, and 6 illustrated the comparison of the classification accuracies using SMOTE (Up sampling) and OVA based on deep learning, stacking, and random forest algorithms, respectively.

From Table 4, the results demonstrated that the proposed method received a higher accuracy rate than other methods when SMOTE and OVA strategy with deep learning as the classifier is applied in almost all datasets except the dermatology dataset. In addition, the SOVA_DL approach obtained the best performance of classification (98.51%) in the new-thyroid dataset.

From Table 5, it was seen that the highest accuracy using SMOTE and OVA with stacking algorithm almost all datasets except dermatology and vertabral3c datasets. The SOVA_SK technique gained the best performance of classification (98.42%) in the pageblocks dataset. In addition, the research outputs indicated that the presented method provided better accuracy values than the single SMOTE approach for the feature with a high imbalance ratio. On the contrary, the datasets, having a low imbalance ratio, outputs were inferior to other methods.

From the results from Table 6, it was seen that the SOVA_RF method is superior to other methods. The SOVA_RF approach achieved the best performance of classification (98.47%) in the dermatology dataset. In addition, the experiment results indicate that the proposed

method offers better classification accuracy values than the single SMOTE approach for the training set with a low imbalance ratio. In contrast, with datasets having a high imbalance ratio, outputs were inferior to other methods.

Table 7, it showed the effectiveness comparison for each approach in terms of Cohen's Kappa. Table 7, when considering the efficiency measured regarding Kappa, showed that the SOVA_RF approach had the highest Kappa (0.981), which was higher than those states of the art models in the dermatology dataset. The SOVA_DL method had a high Kappa (0.974) with the new-thyroid dataset. Furthermore, the SOVA_SK technique had a high Kappa (0.971) in the pageblocks dataset. In summary, a high Kappa coefficient indicated a high agreement of consistent classification, whereas a low value indicated a low agreement of consistent classification.

Table 8 illustrated the performance comparison of model classification in terms of sensitivity and the effectiveness comparison of specificity value was shown in Table 9.

Table 8 indicated that the SOVA_DL method gained sensitivity values more than other methods in 3 datasets; yeast, new-thyroid, and vertabral3c. The outputs from Table 8, indicated that the maximum sensitivity value was 0.9851 using SMOTE and OVA with a deep learning algorithm in the new-thyroid dataset. The SOVA_SK approach achieved sensitivity values better than SOVA_DL and SOVA_RF in the car, pageblocks, and Ecoli datasets. Furthermore, in glass, dermatology, and Cleveland obtained good sensitivity when the SOVA_RF is applied to measure the performance of the classification model.

Table 9 it was seen that the SOVA_RF method received the best specificity values (0.9969). The SOVA_DL technique obtained better specificity than other methods in 3 datasets; yeast, new-thyroid, and vertabral3c. Moreover, the SOVA_SK approach gained the best performance in pageblocks at 0.9960.

The graph compares the accuracy values of the classification model using SOVA_DL, SOVA_SK, and SOVA_RF methods for multiclass imbalanced data handling.
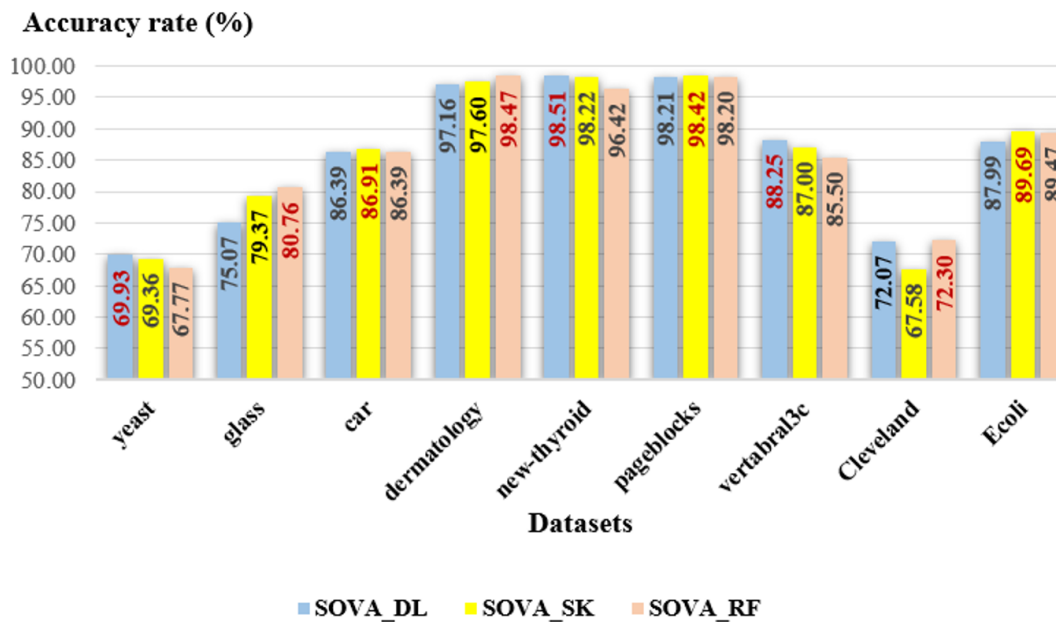

Fig. 3: Comparison of classification accuracies (%) using SOVA_DL, SOVA_SK, and SOVA_RF approaches in multiclass imbalanced datasets.

Table 3: Combining SMOTE approach and one-versus-all strategy with different classifiers used in this study and associated abbreviations

Algorithm	Abbreviation
Combining SMOTE approach and OVA strategy based on deep learning algorithm with 10-fold cross-validation	SOVA_DL
Combining SMOTE approach and OVA strategy based on stacking learning algorithm with 10-fold cross-validation	SOVA_SK
Combining SMOTE approach and OVA strategy based on random forest learning algorithm with 10-fold cross-validation	SOVA_RF

Table 4: The comparison of the classification accuracies (%) using SMOTE (up sampling) and OVA based on a deep learning algorithm for multiclass imbalanced datasets

Dataset	Original training set (No SMOTE)	SMOTE (up sampling)	SOVA_DL*
Yeast	58.83	69.05	69.93
Glass	66.36	74.75	75.07
Car	75.64	86.15	86.39
Dermatology	96.43	97.81	97.16
New-Thyroid	95.78	98.22	98.51
Pageblocks	96.86	98.12	98.21
Vertabral3c	86.13	86.25	88.25
Cleveland	57.94	70.95	72.07
Ecoli	87.49	87.36	87.99

* Proposed method

Table 5: Comparison of classification accuracies (%) using SMOTE (up sampling) and OVA based on stacking algorithm for multiclass imbalanced datasets

Dataset	Original training set (No SMOTE)	SMOTE (up sampling)	SOVA_SK*
Yeast	57.88	69.78	69.36
Glass	65.91	74.01	79.37
Car	74.89	85.97	86.91
Dermatology	96.44	98.68	97.60
New-Thyroid	96.73	97.32	98.22
Pageblocks	96.78	98.16	98.42
Vertabral3c	59.64	86.75	87.00
Cleveland	57.28	70.72	67.58
Ecoli	86.27	89.28	89.69

* Proposed method

Table 6: The comparison of the classification accuracies (%) using SMOTE (Up sampling) and OVA based on a random forest algorithm for multiclass imbalanced datasets

Dataset	Original training set (No SMOTE)	SMOTE (up sampling)	SOVA_RF*
Yeast	53.48	64.37	67.77
Glass	68.23	77.91	80.76
Car	78.07	87.12	86.39
Dermatology	98.09	98.26	98.47
New-Thyroid	93.48	96.11	96.42
Pageblocks	96.00	98.36	98.20
Vertabral3c	82.26	84.50	85.50
Cleveland	56.57	70.28	72.30
Ecoli	87.76	89.69	89.47

* Proposed method

Table 7: The comparison of the classification of Cohen's Kappa using SMOTE (up sampling) and OVA based on deep learning, stacking, and random forest algorithms for multiclass imbalanced datasets

Dataset	SOVA_DL	SOVA_SK	SOVA_RF
Yeast	0.628	0.621	0.598
Glass	0.673	0.730	0.748
Car	0.785	0.796	0.786
Dermatology	0.965	0.970	0.981
New-Thyroid	0.974	0.969	0.938
Pageblocks	0.967	0.971	0.967
Vertabral3c	0.819	0.801	0.775
Cleveland	0.591	0.540	0.585
Ecoli	0.842	0.866	0.862

Table 8: Sensitivity comparison using SOVA_DL, SOVA_SK, and SOVA_RF for multiclass imbalanced datasets

Dataset	SOVA_DL	SOVA_SK	SOVA_RF
Yeast	0.6993	0.6936	0.6780
Glass	0.7563	0.7936	0.8078

Table 8: Continue

Car	0.8639	0.8691	0.8639
Dermatology	0.9716	0.9760	0.9847
New-Thyroid	0.9851	0.9821	0.9642
Pageblocks	0.9821	0.9842	0.9820
Vertabral3c	0.8825	0.8700	0.8775
Cleveland	0.7207	0.6757	0.7230
Ecoli	0.8800	0.8968	0.8947

Table 9: Specificity comparison using SOVA_DL, SOVA_SK, and SOVA_RF for multiclass imbalanced datasets

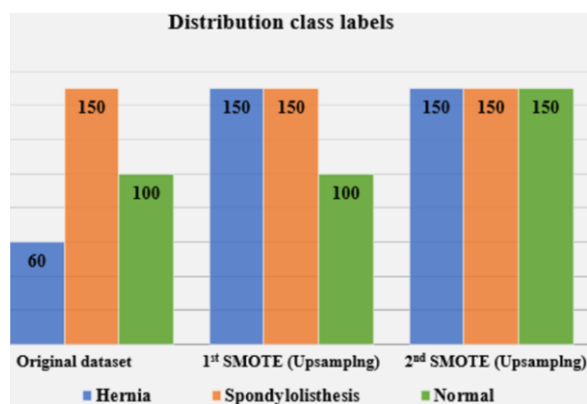| Dataset	SOVA_DL	SOVA_SK	SOVA_RF
Yeast	0.9544	0.9532	0.9498
Glass	0.9353	0.9514	0.9546
Car	0.9501	0.9522	0.9501
Dermatology	0.9942	0.9951	0.9969
New-Thyroid	0.9925	0.9910	0.9766
Pageblocks	0.9955	0.9960	0.9954
Vertabral3c	0.9376	0.9305	0.9218
Cleveland	0.9117	0.8966	0.9126
Ecoli	0.9399	0.9407	0.9403

Fig. 4: The number of instances per class before and after the application of SMOTE (up sampling) in the vertebral 3c dataset

From the bar chart in Fig. 3, it was seen that the results of the SOVA_DL approach were superior to other methods in yeast (69.93%), new-thyroid (98.51%), and vertabral3c (88.25%) datasets whereas in three of the datasets; glass, dermatology, and *Ecoli*, the accuracy rate was lower than the SOVA_SK and SOVA_RF techniques.

On the other hand, the SOVA_RF was a higher performance of classification in three datasets (glass (80.76%), dermatology (98.47%), and Cleveland (72.30%). Furthermore, the good accuracy rates when the SOVA_SK was applied to classify in the car (86.91%), pageblocks, (98.42%), and *Ecoli* (89.69%) datasets.

Overall, it was clear that each method of machine learning improved the performance of model classification for multiclass imbalanced data handling.

The distribution of instances for each class label for the original dataset and after the implementation of SMOTE (up sampling) technique is demonstrated in Fig. 4.

Discussion

The overall results of the proposed method could improve the performance of the classification model for multiclass imbalanced datasets. In this study, the performance was compared with the three classifiers; deep learning, stacking method, and random forest technique.

The results indicated that the SOVA_DL method had better performance than both single SMOTE and without SMOTE methods in all datasets except the dermatology dataset. The outputs of three datasets; yeast, new-thyroid, and vertabral3c offered the best accuracy, sensitivity, and specificity when a deep learning algorithm was used as a classifier. The SOVA_DL had the best performance with the new-thyroid dataset. The classification performances of the SOVA_SK approach were better than those of the other methods in car, pageblocks, and *Ecoli*. In addition, the best result for glass, dermatology, and Cleveland datasets was provided using the SOVA_RF technique, and it is likewise the sensitivity and specificity values of performance classification. The better performance could have been caused by combining the predictions from the multiple classifier's models on the same dataset.

The research also considered the procedural similarity depended on SMOTE and OVA strategy to handle the multiclass imbalanced problem. The difference between the presented method and previous work (Puttiporn and Yaowares, 2019) was the use OVA technique with deep learning and ensemble classifiers; stacking, and random forest algorithms. In this study, OVA was used to split the multiclass training set into multiple binary training sets. The three classifiers using a deep learning algorithm with an H2O operator, stacking algorithm, and random forest algorithm were used to train each binary classification model. In addition, these algorithms were to assess the

efficiency of the model classification with 10-fold cross-validation, whereas the previous study (Puttiporn and Yaowares, 2019) used SMOTE and ADASYN approaches for the imbalanced data sets of elderly's knee osteoarthritis. The experimental outputs illustrated that using the random forest technique with a gain ratio as the splitting criterion to create individual decision trees which could improve the performance of imbalanced classification.

In conclusion, these results indicated that both the SOVA_DL and SOVA_RF approaches might be a good choice for a dataset that had a small number of classes and a low imbalance ratio whereas the SOVA_SK method could handle the multiclass imbalanced dataset. The prediction performance of the SOVA_SK approach was suitable for the domain that had a large number of instances and a high imbalance ratio. Therefore, the proposed method was used to handle imbalanced datasets and to avoid overfitting problems.

Conclusion

The method of handling multiclass imbalanced datasets using SMOTE based on OVA strategy with deep learning and ensemble classifiers (stacking and random forest algorithms) was presented in this study. In the present study, the performances of SOVA_DL, SOVA_SK, and SOVA_RF approaches were compared with that of a single SMOTE and without up sampling technique. The aims of using SOVA_DL, SOVA_SK, and SOVA_RF methods are to handle multiclass imbalanced datasets and to improve the predictive model. The best effectiveness of classification with an accuracy of 98.51% (sensitivity = 0.9851 and specificity = 0.9925) in the new-thyroid dataset when the SOVA_DL approach was used. The SOVA_SK had the highest accuracy rate was 98.42% (sensitivity = 0.9842 and specificity = 0.9960) in pageblocks dataset. Furthermore, the best accuracy of SOVA_RF method was 98.47% (sensitivity = 0.9847 and specificity = 0.9969) in dermatology dataset. The research results indicated that a hybrid of SMOTE and OVA techniques received a better accuracy rate than the single SMOTE approach because, in preprocessing step, the OVA with three different classifiers is applied to evaluate each binary model and to decide the best final output. In addition, the experiment results indicated that the SOVA_SK method can help increase accuracy for high imbalance datasets with a large number of instances, whereas both the SOVA_DL and the SOVA_RF approaches provided good performance for the domain with low imbalance datasets and a small number of instances.

The benefits of the proposed method for this research are the improvement in the efficiency of classification by increasing the accuracy rate.

Furthermore, this method achieved a good alternative for handling multiclass imbalanced datasets. In the future, the sampling method might be applied to real-world problems using a hybrid under-sampling or adaptive sampling other methods.

Acknowledgment

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the authors have read and approved the manuscript and that no ethical issues are involved.

References

Abd Elrahman, S. M., & Abraham, A. (2013). A review of the class imbalance problem. *Journal of Network and Innovative Computing*, *1*(2013), 332-340.

Almayyan, W. (2021). Mining Sports Articles using Cuckoo Search and Tabu Search with SMOTE Preprocessing Technique. DOI.org/10.3844/jcssp.2021.231.241

Bhagat, R. C., & Patil, S. S. (2015, June). Enhanced SMOTE algorithm for classification of imbalanced big data using random forest. In the *2015 IEEE International Advance Computing Conference (IACC)* (pp. 403-408). IEEE. DOI.org/10.1109/IACC35547.2015

Bisht, N., Ahmad, A., & Bisht, S. (2016). Application of feature selection methods and ensembles on network security dataset. *International Journal of Computer Applications*, *135*(11), 1-5.

Bolon-Canedo, V., Sanchez-Marono, N., & Alonso-Betanzos, A. (2011). Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset. *Expert Systems with Applications*, *38*(5), 5947-5957. DOI.org/10.1016/J.ESWA.2010.11.028

Candel, A. & LeDell, E. (2022). H2O's Deep Learning Architecture. In Deep Learning with H2O, 6, A. Bart, (Eds.), *Mountain View*, CA, USA, pp: 10–21.

Charte, F., Rivera, A., Jesus, M. J. D., & Herrera, F. (2013, September). A first approach to deal with the imbalance in multi-label datasets. In *International conference on hybrid artificial intelligence systems* (pp. 150-160). Springer, Berlin, Heidelberg. DOI.org/10.1007/978-3-642-40846-5_16

De Raadt, A., Warrens, M. J., Bosker, R. J., & Kiers, H. A. (2019). Kappa coefficients for missing data. *Educational and Psychological Measurement*, *79*(3), 558-576. DOI.org/10.1177/0013164418823249

Farid, D. M., Zhang, L., Rahman, C. M., Hossain, M. A., & Strachan, R. (2014). Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. *Expert systems with applications*, *41*(4), 1937-1946. DOI.org/10.1016/J.ESWA.2013.08.089

Ghanem, A. S., Venkatesh, S., & West, G. (2010, August). Multi-class pattern classification in imbalanced data. In 2010 20th *International Conference on Pattern Recognition* (pp. 2881-2884). IEEE. DOI.org/10.1109/ICPR.2010.706

Hasib, K. M., Iqbal, M., Shah, F. M., Mahmud, J. A., Popel, M. H., Showrov, M., ... & Rahman, O. (2020). A survey of methods for managing the classification and solution of the data imbalance problem. *arXiv preprint arXiv:2012.11870* DOI.org/10.3844/jcssp.2020.1546.1557

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 1-54. DOI.org/10.1186/s40537-019-0192-5

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing the high-class imbalance in big data. *Journal of Big Data*, *5*(1), 1-30. DOI.org/10.1186/s40537-018-0151-6

Mahani, A., & Ali, A. R. B. (2019). Classification problem in imbalanced datasets. *Recent Trends in Computational Intelligence*, 1-23.

Maheshwari, S., Jain, R. C., & Jadon, R. S. (2018). An Insight into Rare Class Problem: Analysis and Potential Solutions. *J. Comput. Sci.*, *14*(6), 777-792.

Majumder, A., Dutta, S., Kumar, S., & Behera, L. (2020). A Method for Handling Multi-class Imbalanced Data by Geometry-based Information Sampling and Class Prioritized Synthetic Data Generation (GICaPS). *arXiv preprint arXiv:2010.05155*. DOI.org/10.48550/arxiv.2010.05155

Mehra, N. & Gupta, S. (2013). Survey on Multiclass Classification Methods. *International Journal of Computer Science and Information Technologies*, 4(4): 572 – 576.

Özdemir, A., Polat, K., & Alhudhaif, A. (2021). Classification of imbalanced hyperspectral images using SMOTE-based deep learning methods. *Expert Systems with Applications*, *178*, 114986. DOI.org/10.1016/j.eswa.2021.114986

Panthong, R., & Srivihok, A. (2019). Liver cancer classification model using hybrid feature selection based on a class-dependent technique for the central region of Thailand. *Information*, *10*(6), 187. DOI.org/10.3390/info10060187

Putri, V. M., Masjkur, M., & Suhaeni, C. (2021, March). Performance of SMOTE in a random forest and naive Bayes classifier for imbalanced Hepatitis-B vaccination status. In *Journal of Physics: Conference Series* (Vol. 1863, No. 1, p. 012073). IOP Publishing. DOI.org/10.1088/1742-6596/1863/1/012073

Puttiporn, T., & Yaowares, S. (2019). A data classification technique was developed for an imbalanced dataset of knee osteoarthritis in the elderly. *Thai Science and Technology Journal*, 1164-1178.

Sáez, J. A., Krawczyk, B., & Woźniak, M. (2016). Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. *Pattern Recognition*, *57*, 164-178. DOI.org/10.1016/j.patcog.2016.03.012

Sun, Z., Song, Q., Zhu, X., Sun, H., Xu, B., & Zhou, Y. (2015). A novel ensemble method for classifying imbalanced data. *Pattern Recognition*, *48*(5), 1623-1637. DOI.org/10.1016/j.patcog.2014.11.014

Upadhyay, K., Kaur, P. & Prasad, S. (2021). A Review on Data level Approaches to address the Class Imbalance Problem. *Proceedings of International Conference on Recent Challenges in Engineering Science Technology*, April 9-10, India, pp: 151–158.

Waqar, M., Dawood, H., Dawood, H., Majeed, N., Banjar, A., & Alharbey, R. (2021). An efficient SMOTE-based deep learning model for heart attack prediction. *Scientific Programming*, *2021*. DOI.org/10.1155/2021/6621622

Xie, W., Wei, S., Zheng, Z., Chang, Z., & Yang, D. (2022). Developing a stacked ensemble model for predicting the mass of fresh carrots. *Postharvest Biology and Technology*, *186*, 111848. DOI.org/10.1016/j.postharvbio.2022.111848

Yu, L., & Zhou, N. (2021). Survey of imbalanced data methodologies. *arXiv preprint arXiv:2104.02240*.

Yuan, X., Xie, L., & Abouelenien, M. (2018). A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data. *Pattern Recognition*, *77*, 160-172. DOI.org/10.1016/j.patcog.2017.12.017