Hybrid SMOTE and Bootstrap Sampling for Imbalanced Classification in Elderly Health Condition Dataset

Rattanawadee Panthong*
University of Phayao, Phayao, THAILAND

Abstract

The dataset was applied in the analysis of the real-world problem which found that the data lacked balance and had several classes due to the data nature and the data collection limitation. This made the data gained a high term of possible imbalance. When the imbalanced data is used in learning, it may reduce the efficiency of data classification. Thus, this study aimed to manage the imbalanced data via hybrid data-level technique using SMOTE and bootstrap approaches based on one versus all with different learning. The four different learning methods; deep learning, stacking algorithm, random forest, and gradient boosting tree are applied to improve the accuracy rate of the classification model. The elderly health condition dataset was obtained from the Meaka Heath Promotion Hospital of Phayao in Thailand. The experimental results indicated that the HySM_BT50% method gained the highest correctness value at 90.11% (Sensitivity = 0.8469, Specificity = 0.9749, and G-means = 0.9514) when using random forest algorithm as a classifier.

Key Words: SMOTE; bootstrap; imbalance; multiclass; classification; deep learning; ensemble method; elderly health condition.

1 Introduction

Most problems found in machine learning are the imbalanced data and have a lot of multi-classes from the dataset occurring from the distribution of the sample group or unequal label class, for example, one class has the least proportion compared with other classes resulting in predicting or classifying the minority class with having low efficiency and misclassification [2, 3]. This is because the data classification method will give good efficiency when the data is balanced or close to each other. The imbalanced data occurs in different domains such as the problem of the medical cluster classification (cancer data and no cancer patients), the problems of the risk management and the anomaly detection [6]. It is quite difficult to gain the data in each cluster having an equal number.

In real-world problems, the most prevalent problem found is that the multi-class imbalanced data set used the form of the text or the image data. A multi-class problem is the classification problem where the instances of data classification fall into one or more than two classes. Thus, the problem of data classification is a challenging issue for the management of the imbalanced multi-class data. The multi-class classification becomes the main problem of machine learning due to the data having several label classes leading to processing the data more complicated. Moreover, in classifying the multi-class data, it shows that the model efficiency in classifying data relies on the majority vote and the prediction of the new class data [24]. The most common solution, involving multi-class problems which are likely to be harder to predict, is to transform them into several binary problems. One-Versus-All (OVA) decomposes classes into binary and then learners improve the representation of the minority examples. This method is easy and useful for the work requiring to classify the multi-class which focuses on adjusting the data to be appropriate for an algorithm [15, 24].

In general, two data-level approaches used in classifying the imbalanced data are applied to increase the samplings in minority data by over sampling and to decrease the samplings in majority data by under sampling. The technique, most likely used by many experts, is SMOTE (synthetic minority over sampling technique) which is introduced to manage the imbalanced data [12, 19, 23, 26]. It is a way to increase the data in the minority group resulting in spreading the more balanced data clusters in order to make the classification of the minority group much better and the model classification much more accurate. The under sampling approach is a way of reducing the majority sample group to have the same or near amount of the minority sample group leading to adjusting the dataset balance before applied for the training model.

The survey data of the elderly health condition gained from Maeka Health Promotion Hospital, Phayao Province in the total of 1194 instances is divided into 3 classes; class 0 is a group of 610 social bound instances, class 1 is a group of 529 home bound instances and class 2 is a group of 55 bed bound instances. From the data of the elderly health condition, it shows the imbalanced data normally found that a number of healthy people is higher than unhealthy people. In assessing and planning the elderly health care, the minority data is more interesting than the majority data which means that the total of instances of class 0 and class 1 put together is higher than class 2. This will result in classifying the data incorrectly or putting in a wrong cluster and making the classification of the minority group having low efficiency. The adjustment of the imbalanced

*Department of Information Technology, School of Information and Communication Technology, , rattanawadee.pa@up.ac.th, Tel.: +66-89-8104181

data of class 2 can be done by data level approach.

Therefore, hybrid data level approaches (SMOTE and bootstrap sampling) based on OVA technique with different learning methods empirically are applied to handle the imbalanced multi-class dataset. In this research, the synthesis of the new sample cluster for a small class (SMOTE) is another option or way applied to manage the classification of the imbalanced data to improve the model efficiency in predicting the data more accurately. Under sampling is the sampling technique by reducing the number of instances, which the majority class has the same amount as the minority class [32]. In this work, bootstrap technique is used as an under-sampling method. Bootstrapping approach is used to decrease the instance of the majority class group. In addition, the OVA strategy with four different classifiers (deep learning, stacking algorithm, random forest, and gradient boosting tree) is trained to evaluate and classify the models. The OVA strategy is applied to change the multi-class learning problem into a two-class learning one. The benefits of the presented method is the improvement in the efficiency of the model classification. This method can help improve the model efficiency in classifying the data more accurately in the group having a small number of samples. Additionally, it provides good efficiency for the domain with the imbalanced multi-class.

2 Background and Related Work

2.1 Elderly Health Condition

The elderly groups are people over 60 years and divided into three groups in Thailand; social bound, home bound and bed bound. Group 1- Social bound refers to the elderly who can help themselves and lead normal lives independently. Moreover, they are able to do their basic routines continuously, being in good health and having no chronic diseases or no more than two chronic diseases which can be controlled. Furthermore, they can help other people, society, community and can participate in social activities. Group 2- Home bound refers to the elderly who can help themselves, still need help from other people in some instances, have some limitations in leading their lives, have chronic diseases which cannot be cured and have both physical and mental complications resulting in doing their basic routines. Group 3- Bed bound refers to the elderly who cannot help themselves do their basic routines completely, need other people to move them, have chronic diseases which cannot be controlled and have complications, not able to help themselves or are paralyzed [27].

2.2 Imbalanced Data Problem

The main problem of the data classification is the imbalanced dataset which is caused by having two data clusters or more meaning that the data in each class is not equal. Definitely, if the data with clearly different quantity is taken through the classification method, the learning model will be classified in a majority group. There are several solutions for managing imbalanced data problems such as data level approaches, algorithm level approaches, and cost-sensitive approaches [11, 28].

2.3 Data-Level Approaches

Data-level is a technique to increase or decrease instances from the imbalanced datasets which can improve the classification accuracy. Data-level approaches play a vital role in imbalanced classification by reducing the distribution of examples and adjusting the balance of classes. These techniques are used to prepare a dataset before the classification stage. After applying datalevel solutions, the training set can learn data more efficiently [2]. Data-level solutions are separated into three categories; oversampling, undersampling and hybrid methods [11,2 8].

The oversampling technique or upsampling increase numbers of the minority class data to have a nearly equal amount of data to the majority class by using SMOTE technique [12, 19, 23, 26, 32]. SMOTE is a technique used in solving the classification of the imbalanced data because the data in each class has a different amount causing the data classification results to increase the minority class and lead to make data have more balance [14, 31]. In the random step, only one value is taken from the data values in the minority class to increase the amount of the minority class which makes the data set have more balance. In the data value random in minority class, one sample value is taken, followed by considering the K-nearest data value. Next, the Euclidean distance is calculated between the randomized data value and each nearest data value to find the least distance value. After that, the new data is synthesized to have the same value with the data giving the least distant resulting in making the amount of the data balance in every cluster.

Under sampling or down sampling is a way of reducing the majority sample group to have the same or nearly the amount of the minority sample group leading to adjusting the dataset balance before applying the training model [32]. Bootstrap sampling is a statistical technique used in resampling with replacement from the original sample. The principle of sampling technique is to increase the amount of data by sampling several clusters where each cluster will have two out of three of the previous data size. By this method, each cluster will not have the same data 100 percent of the time resulting in gaining several sampling clusters from the previous data. In this study, the bootstrap sample is used to randomize the sample only one value at a time with the total amount of n times. The values gained will return to the dataset before conducting the next sampling [5, 35].

The hybrid data level approach is the combination between oversampling and under sampling [29, 32]. It is used to increase and decease the instances of each class before the classification step. Moreover, it is the way to take the over and the under samplings to be applied to find the middle value in sampling the dataset between the two groups. This is because increasing too much data may result in causing the data bias while decreasing too much data may lose the important data in model creating. Hence, this method can help handle the imbalanced multi-class.

2.4 Multi-Class Classification.

The multi-class classification method is the way to classify the data set of more than two classes via changing the problem and dividing the classification into two types. Generally, the popular method of classifying the multi-class into two classes is the OVA approach which is the method of reducing the learning technique from the multi-class learning problem to the two-class learning problem. The OVA method is classified as a K- binary classifier. The classier is built for each class so the classifier is trained by the C_i class and the total instance groups of other classes. The result from the binary classification can make a decision by using this function: $F(x)$ = argmaxi=1…K $f_i(x)$ determining to choose the highest value with the sample tested by OVA method; K-binary classifier. The classifier is built for each class so the classifier is trained by the C_i class and the all-sample group of other classes [7]. The advantage of OVA strategy is the fast-processing time consumption. Thus, this study chooses the OVA technique applied with different learning methods to reduce the processing complexity of the classification model and improve the model efficiency in classifying the data to be more accurate [15, 18].

2.5 Ensemble Method

Ensemble method is the technique focusing on improving the accuracy of the simulated model result using several classification models to help find the answer of which the combination of the models can increase the high accuracy of the results. This method tends to take only one training dataset [16]. However, to make the model have more functions, the model is created by using different classification techniques. After creating the ensemble models, the models are taken to predict the new data. Due to the ensemble model having several models, each model will give its result and all results will be considered to have the most appropriate answers through voting [17, 34].

2.6 The Performance Assessments of Classifiers are as Followed

Accuracy is one metric measurement for classification models.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Where TP = the number of true positives, TN = the number of true negatives, FP = the number of false positives, and FN = the number of false negatives [15].

Sensitivity or true positive rate is the proportion of instances where it is predicted as a positive label [15].

$$SN = \frac{TP}{TP+FN} \quad (2)$$

Specificity or true negative rate is the proportion of the instances where it is predicted as a negative label [15].

$$SP = \frac{TN}{TN+FP} \quad (3)$$

G-means (geometric mean) is the measurement performance of the class-imbalanced classifiers. It is applied to balance the sensitivity and the specificity values. If the sensitivity and the specificity are equal, G-means has the maximum value [15].

$$G - mean = \sqrt{(SN * SP)} \quad (4)$$

2.7 Related Work

Other researchers have presented a combination of data level techniques using SMOTE and under sampling. Pristyanto, et al., [21] presented the data level approach to balance the class distribution on educational data mining (EDM). The combination of the SMOTE and the OSS techniques were used to handle the imbalance on educational data mining. The OSS approach is applied to split the majority class into four sections; noise sample, borderline sample, redundancy sample, and safety sample. The results indicated that the hybrid of SMOTE and One-Sided Selection (OSS) provided the best effectiveness when using SVM as a classifier.

The hybrid simulated annealing approach (SA) with different classifiers (discriminant analysis, SVM, decision tree, and KNN) was presented by Desuky et al., [5]. The SA method is applied to manage the unbalanced data and to select majority examples on UCI and KEEL datasets. Furthermore, the F-score metric with objective function is used to choose instances. The outputs of experiment illustrated that the SA method with 4 different classifiers received the best accurate rate on 9 binary datasets.

Kou et al., [13] combined the methods of resampling and ensemble model for the credit and the finance evaluation data. In this work, the resampling technique was applied to the clustering and distance-based imbalance learning mode or called the CDEILM. Moreover, the cluster-size-based resampling model was developed to divide the group size of the under sampling rate in the resampling step. The outputs of a hybrid approach illustrated that the effectiveness of AUC and G-measure values were higher than other approaches.

In addition, this method could help solve the problems of the domains in the finance imbalanced datasets.

Zhaozhao et al., [33] presented the method of the hybrid SMOTE and k-means for the imbalanced medical data. The cluster-based oversampling algorithm was applied to identify the class of all instances. The results of this paper indicated that the sensitivity and the specificity values were at 99.84% and at 99.56% when the random forest algorithm is used to classify the model. Astha, et al., [1] proposed the method of the SMOTE and the cluster-based undersampling for the imbalanced multiclass. This approach was applied to balance and preprocess the training sets. The SCUT method received higher accuracy than other methods when applied for the preprocessing data.

The combination SMOTE and undersampling method for the imbalanced datasets was proposed by Hanskunatai [9]. The

hybrid sampling technique between SMOTE and undersampling technique by using the decision tree and naïve bayes was the data classifier. The DBSCAN algorithm is used to divide a group of positive and negative classes. Negative class is removed 50% with this algorithm. The results demonstrated that the hybrid sampling technique obtained the F-measure more than other sampling approaches in 7 datasets (Haberman, Glass6, glass0, vehicle2, new-thyroid1, new-thyroid2, and yeast3) when decision tree was used as the classifier. The outputs indicated this method could help improve performance of predictive model. Furthermore, the proposed method achieved the highest F-measure in Wiscosin at 0.962. Additionally, Xu et al., [23] presented a technique of the new hybrid SMOTE, Tomek-link and combined cleaning and resampling (CCR-SMOTETL). In this work, the CCR-SMOTETL approach was used to random the instances and to detect the noise data. The research results demonstrated that the effectiveness of the classification received better accuracy than other methods when classified using Random Forest algorithm.

3 Materials and Methods

3.1 Dataset

The dataset of the elderly health condition survey used in this experiment was received from Maeka Health Care Promotion Hospital, Muang District, Phayao Province, Thailand. The data was collected from January, 2022 to February 2022 in which the details of the dataset is shown in Table 1. The elderly health condition dataset is divided into three different classes; social bound (class 0), home bound (class 1), and bed bound (class 2). An example of the dataset for the experiment is shown in Figure 1. The descriptions of each feature for the elderly health condition data are presented in Table 2 in supplementary materials. Furthermore, the names of the data level approach and the associated abbreviations are shown in Table 3.

Table 1: Showing the details of the data feature used in the experiment

Dataset	No. of features	No. of instances	No. of instances/ Class	Imbalanced ratio (IR)
Elderly health condition	34	1194	610/529/55	11.09

Table 2: The explanation of the value substitution of data in classification

Feature	Descriptions	Feature	Descriptions
1. Occ (Occupation)	1= Agriculture, 2=General employee, 3=Merchant/ Personal business, 4=Animal keeper, 5= Government pensioner, 6=Unemployment, 7= Others	2. Marital Status	1= Single 2= Married 3= Widowed 4=Divorced
3. BP1 (Blood pressure check, first time)	1=normal (not more than 120/80), 2=starting high (120-139/80-89), 3=high (140-159/90-99), 4=very high (>160/100) 0= no information	4. BP2 (Blood pressure check, second time)	1=normal (not more than 120/80), 2=starting high (120-139/80-89), 3=high (140-159/90-99), 4=very high (>160/100) 0= no information
5. Cdis (Having chronic disease)	1=yes, 2=No	6. Phydis (Physical disability)	1=yes, 2=No
7. Mact (Moderate movement activity)	1=Having regularly at least 3 times a week 2=Having irregularly less than 3 times a week 3= No having due to disability	8. Teeth (The present having good teeth at least 20 teeth)	1=yes, 0= No
9. Smoking	1= Never smoking, 2= Ever smoking but not smoking now, 3= Smoking	10. Alcohol Drinking	1 = Never drinking 2= Ever drinking but stopping now 3= Still drinking
11. Eye sight problem	1=yes, 0= No	12. Eye Results (The result of eye sight examination)	1=short-sighted eye, 2=long-sighted eye, 3=cataract, 4=glaucoma, 5= pterygium, 6= macular degeneration

13. Feeling_ unpleasant	1=yes, 0= No	14. Feeling _depressed	1=yes, 0= No
15. Assess_fall (The assessment of falling condition)	1= less than 30 seconds 2=more than 30 seconds 3= unable to walk	16. Urinary (The screening of urinary incontinence)	1=yes, 0= No
17. Assess remem1 (The assessment of the dementia condition by remembering and telling all words of things correctly.	1= true 2= false	18. Assess_remem2 (The assessment of the dementia condition by telling one's own name and age correctly)	1=true 2= false
19. Results_of yearly diabetes check up	1=normal 2= abnormal 3= not checking	20. Results of yearly hypertension check up	1=normal 2= normal 3= not checking
21. Assess_sleep (Sleep problem diagnosis)	1=Have, 0= No have	22. Assess_knee (Knee osteoarthritis diagnosis)	1=yes, 0= No
23. ADL1 (Feeding)	0= unable, 1= needs help cutting, spreading butter 2=independent (food provided within reach)	24. ADL2 (Grooming)	0 = needs help with personal care 1 = independent face/hair/teeth/shaving (implements provided)
25. ADL3 (Transfer)	0 = unable – no sitting balance 1 = major help (one or two people, physical), can sit 2 = minor help (verbal or physical) 3 = independent	26. ADL4 (Toilet use)	0 = dependent 1 = needs some help, but can do something alone 2 = independent (on and off, dressing, wiping)
27. ADL5 (Mobility)	0 = immobile 1=wheelchair independent, including corners 2 = walks with help of one person (verbal or physical) 3 = independent	28. ADL6 (Dressing)	0 = dependent 1 = needs help, but can do about half unaided 2 = independent (including buttons, zips, laces)
29. ADL7 (Stairs)	0 = unable 1 = needs help (verbal, physical, carrying aid) 2 = independent up and down	30. ADL8 (Bathing)	0 = dependent, 1 = independent (or in shower)
31. ADL9 (Bowels)	0 = Incontinent (or needs to be given enema) 1 = occasional accident (once/week) 2 = continent	32. ADL10 (Bladder)	0 = incontinent, or catheterized and unable to manage 1 = occasional accident (max. once per 24 hours) 2 = continent (for over 7 days)
33. BMI (body mass index (nutritional status))	1=obese range 2=healthy weight range 3=overweight range 4=obese range 5=very obese range 0=Unknown	34. Class (Group of elderly which divided by their abilities in doing their daily lives)	Class0= Social Bound Class1= Home Bound Class2= Bed Bound

	ADL6	ADL7	ADL8	ADL9	ADL10	Results_d	Results_h	Teeth	Results_e	Results_e	Feel_dep	Feel_bore	Assess_re	Assess_re	Assess_fa	Urinary	Assess_sl	Assess_kn	Body_mas	Class
2	2	2	1	2	2	3	1	1	2	2	1	1	1	1	1	2	1	2	2	class0
3	2	2	1	2	2	1	1	2	2	3	2	2	1	1	1	1	2	1	1	class0
4	2	2	1	1	1	1	1	2	2	3	1	1	1	1	2	1	1	1	1	class0
5	2	2	1	2	2	1	1	1	1	2	2	2	1	1	2	2	1	1	2	class0
6	2	2	1	2	2	1	1	1	1	0	2	2	1	1	2	2	1	1	2	class1
7	2	2	1	2	2	1	1	1	2	2	1	1	1	1	1	2	2	1	2	class0
8	2	2	1	2	2	1	1	2	2	2	2	2	1	1	2	2	2	1	3	class0
9	2	2	1	2	2	1	1	2	2	1	2	2	1	1	2	2	2	1	2	class1
10	2	2	1	2	2	2	2	1	2	1	2	2	1	1	1	2	2	1	3	class2
11	2	2	1	2	2	1	1	1	1	0	2	2	1	1	2	2	2	2	2	class2
12	2	2	1	2	2	3	1	1	1	0	2	2	1	1	1	2	2	1	2	class2
13	2	2	1	2	2	3	1	1	1	0	2	2	1	1	2	2	2	1	2	class1
14	2	2	1	2	2	1	1	1	1	0	2	2	1	1	1	2	2	2	2	class1
15	2	2	1	2	2	1	1	1	1	0	2	2	1	1	1	2	2	1	4	class1
16	2	2	1	2	2	1	1	2	1	0	2	2	1	1	2	2	2	2	4	class1
17	2	2	1	2	2	1	2	2	1	0	2	1	1	1	1	2	1	1	0	class1
18	2	2	1	2	2	1	1	2	2	7	2	2	2	2	2	2	1	1	0	class0
19	2	2	1	2	2	1	1	1	2	2	1	1	1	1	2	1	1	1	3	class0
20	2	2	1	2	2	1	1	1	2	2	2	2	1	1	1	2	1	1	5	class0

Figure 1: The example of the elderly health condition dataset

Table 3: Data level approach used in the present research and the associated abbreviations

Algorithm	Abbreviation
SMOTE based on OVA strategy	SMOTE_OVA
Hybrid SMOTE and bootstrap sampling based on OVA with different learning methods (the size of the examples is 40% of majority class)	HySM_BT40%*
Hybrid SMOTE and bootstrap sampling based on OVA with different learning methods (the size of the examples is 50% of majority class)	HySM_BT50%*
Hybrid SMOTE and bootstrap sampling based on OVA with different learning methods (the size of the examples is 60% of majority class)	HySM_BT60%*

* The proposed method

3.2 Proposed Method

In this study, the researcher presents the method of the imbalance multiclass management by applying SMOTE and bootstrap techniques based on an ensemble learning method shown in the research conceptual framework as in Figure 2. The processes of the proposed method are as follows.

Step 1: Data collection

This part involves data collection on the elderly's health condition survey. Then, the data is kept in a MySQL database and changed into a needed file, a CSV file, easy to be applied.

Step 2: Data preparation

The method of the missing data substitution has many techniques which are statistical method or mining technique used for substituting the missing data, for example, listwise data deletion [10] is an easy technique in managing the missing data. This approach is done by analyzing only the complete data. It is suitable when having a small amount of missing data and the synthesis result is very clear which is mainly applied by default. For presentation, the substitution technique of missing data is used with the unknown value.

Mean substitution is a technique of replacing the missing data by the value-known data mean in each small group of the variables [10]. It is used due to the hypothesis that the value of the missing data should be relied on the sample unit feature which should have the same interesting data value. Moreover, the class instances of unknown values will be deleted. In this case, the data has many lost features, but not more than 5 percent of all data through deleting the instances having missing data.

Data transformation is the method of converting, extracting and mapping data into a usable format. In this research, the discretization technique is used to convert numerical to nominal data. Discretization is a method of changing the number data value into the data with a small size data. This method is done before the data preparation step which reduces the data process through decreasing the size and the data complexity [8]. The steps of discretization value used in clustering the features is determined by the user and the expert. In addition, the attributes are discretized by claiming the values from the laboratory.

Step 3: Hybrid data level approaches using SMOTE and bootstrap sampling technique

Data preprocessing is to prepare the data processing and managing the data before classification. It is the data preparation which is an important process of the machine learning. If the data preparation is not done well, it will result in the operating efficiency of the other processes. In this step, the problem of the dataset having different spreading classes is managed by adjusting the class balance to have close or an equal number through the combination of SMOTE and bootstrap techniques in solving the imbalanced problems. SMOTE helps synthesize the minority class to have the same size as the majority class while the undersampling approach reduces the sample group with the majority class having the same number

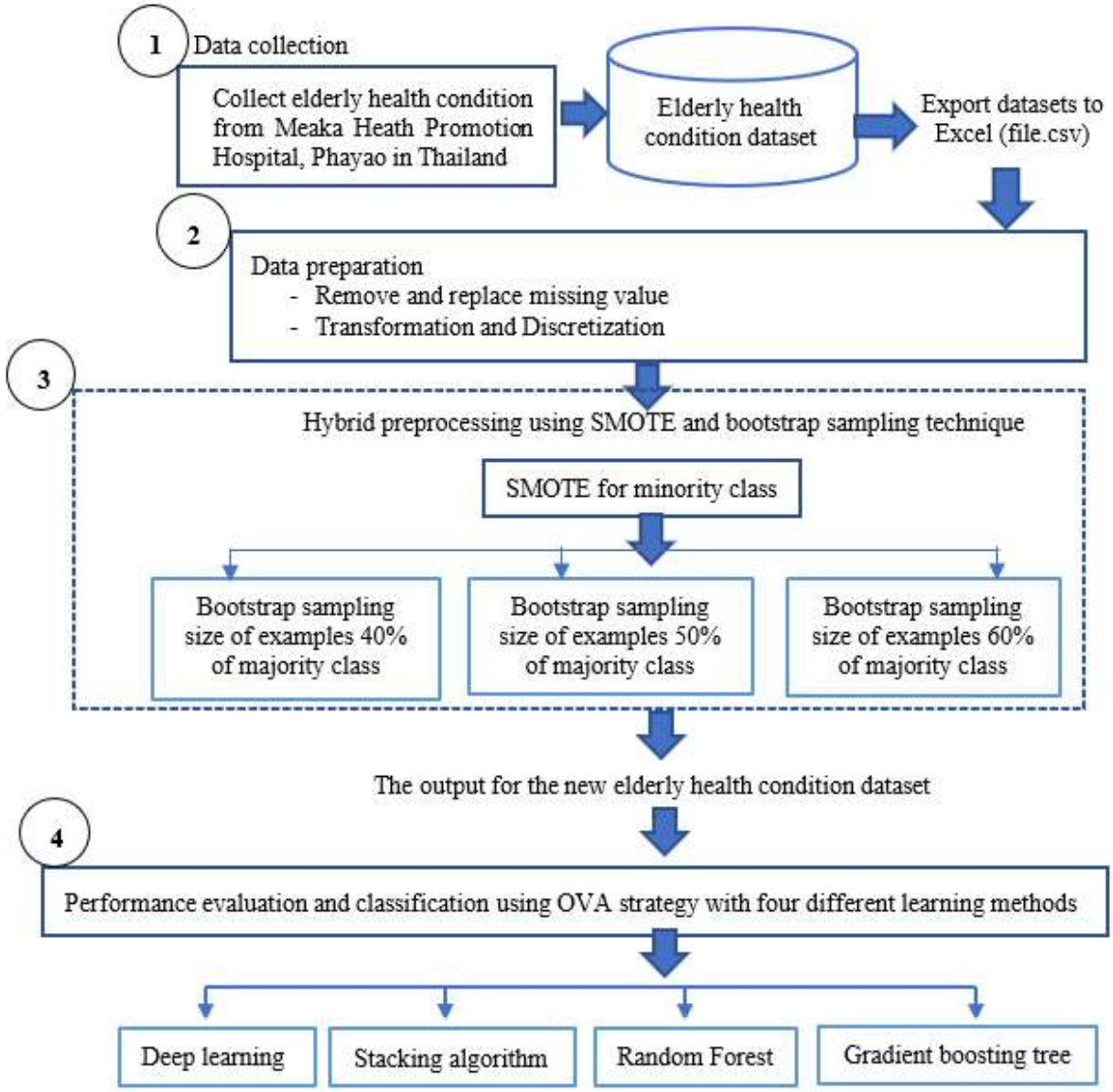

Figure 2: Framework of hybrid data level approaches based on stacking learning for imbalanced classification

or size as the minority class sample. In this research, bootstrap sampling is used in randomizing the data sample in order to make each class balanced before creating the data classification. The bootstrap sampling uses the operator to select the instances according to the sample size. The best configuration of parameter sample size is found by optimizing the parameters process using the RapidMiner software with approximately 40% - 60% of the majority class.

Step 4: Performance evaluation using deep learning and ensemble classifiers for multi-class classification

This research has improved the efficiency of the multi-class classification by using OVA strategy to divide the multi-class problem into two class or binary class together. After adjusting the dataset via the combination SMOTE and bootstrap technique, the new dataset is classified into a cluster by using OVA technique to define the main class as a positive class and the rest of other classes to be a negative class. Then, the new dataset is led to create a model to classify the data type by applying four learning classifiers. In this step, four learning methods; deep learning, stacking algorithm, random forest, and gradient boosting tree are used for the classification and evaluation models [30].

- Deep learning (DL) [20, 22, 25] is a category of neural network model as a multilayer perceptron (MLP). It is used to train with learning algorithms. Deep learning as supervised learning can help seek patterns from a large data and create models for model prediction.
- Stacking algorithm (SK) is a type of the ensemble learning method. It is one of the most popular ensemble classifiers.

This method typically creates a heterogeneous ensemble. It can combine the predictions from multiple classifies. Stacking method is used to train multiple models for solving similar problems [25, 34].

- Random forest method (RF) [4, 34] is an ensemble decision tree. By the principle, the random forest trains the same model for several times or instances on the same data set. Each time the training will choose a different trained data. Then all model decision making is voted upon. The advantage of this algorithm is to consume the rapid time in training the model. Furthermore, it avoids the risk of overfitting.
- Gradient boosting tree (GBT) is an ensemble method that helps improve the accuracy of trees. It can train either regression or classification tree models. The GBT learning algorithm is similar to a boosting technique and a decision tree. The principle of GBT is that the sampling and creating of the decision tree from different simulations and each simulation is assessed until gaining the complete decision tree model. The principle of the GBT method is to create multiple decision tree models and to evaluate each model of a decision tree. This algorithm provides a complete decision tree [4].

4 Results

In this section we will present the results of the experiment and discussion through comparing combinations of SMOTE and undersampling technique based on different classifiers (deep leaning, stacking algorithm, random forest, and gradient boosting tree) and single sampling techniques. It is compared with the accuracy rate, sensitivity, specificity, and G-means in the elderly health condition dataset.

For Table 4, it is the comparison of the effectiveness of accuracy on classification data via hybrid data level approaches; SMOTE and bootstrap technique, based on OVA strategy with different learning methods. From Table 4, it shows the effectiveness of the model classification through the HySM_BT50% method by using RF algorithm gaining the highest accuracy value at 90.11% as well as the HySM_BT60% when using RF method as classifiers at 88.48%. For the HySM_BT40% technique using SK algorithm has good accuracy of classification at 88.13%. In Table 4, it is clearly seen that the output of the HySM_BT50% method obtained is the best at an average accuracy rate of 88.01%.

Moreover, the HySM_BT50% using DL method is equal to that of the HySM_BT50% by GBT algorithm (87.18%). On the other hand, the SMOTE_OVA technique learns better than HySM_BT40% and HySM_BT60% approaches when using GBT as classifier. In summary, the performance of classification is superior to the other methods when the imbalanced data management is applied by using the hybrid data level approach based on OVA strategy with four different classifiers.

Tables 5 and 6, show the comparison of the classification efficiency in terms of sensitivity and specificity values. From Table 5, the result indicates the highest sensitivity gain is at 0.9011 when combining SMOTE and bootstrap sampling technique based on OVA with RF classifiers. Also, in Table 6, the specificity value is classified by RF algorithm which gains the highest value at 0.9480.

The result from Table 7 indicates that the HySM_BT50% by using RF classifier to assess the model efficiency which shows the G-means value gaining higher than other methods. Likewise, the HySM_BT40% with SK and HySM_BT60% with RF algorithm obtains the high G-means. The HySM_BT50% method gives the best G-means value at 0.9242.

The implementation of hybrid data level methods uses SMOTE and bootstrap technique. The chart in Figure 3 compares the distribution of samples of each class after resampling data. There are 3 methods consisting of HySM_BT40%, HySM_BT50%, and HySM_BT60%. Each class represents a group of the elderly. Class_High or class 0 refers to a group of the elderly which are social bound; Class_Middle or class 1 refers to a group of elderly that are home bound; and Class_low or class 2 refers to a group of elderly that are bed bound.

5 Discussion

When considering the efficiency of the hybrid sampling technique together with one versus group, it shows that the technique presented gives the highest correctness value when classifying the data by ensemble learning (RF). For example, the HySM_BT50% method achieves the maximum average accuracy using four different learning methods (DL, SK, RF,

Table 4: Performance of classification accuracy (%) using hybrid data level approaches based on four different learning methods

Method	No. of Instances	DL	SK	RF	GBT	Average
Original dataset	1194	71.51	72.91	75.14	72.35	72.98
SMOTE_OVA	1830	85.06	84.52	84.88	85.79	85.06
Bootstrap	165	83.33	83.33	81.25	77.08	81.25
HySM_BT40%*	732	85.39	**88.13**	87.67	84.47	86.42
HySM_BT50%*	915	**87.18**	87.55	**90.11**	**87.18**	**88.01**
HySM_BT60%*	1098	85.15	84.24	88.48	83.94	85.45

Table 5: Comparison of sensitivity using hybrid data level approaches based on four different learning methods

Method	DL	SK	RF	GBT
Original dataset	0.7151	0.7291	0.7514	0.7235
SMOTE_OVA	0.8506	0.8452	0.8488	0.8579
Bootstrap	0.8333	0.8333	0.8125	0.7708
HySM_BT40%*	0.8539	0.8813	0.8767	0.8447
HySM_BT50%*	0.8718	0.8755	0.9011	0.8718
HySM_BT60%*	0.8515	0.8424	0.8848	0.8394

Table 6: Comparison of specificity using hybrid data level approaches based on four different learning methods

Method	DL	SK	RF	GBT
Original dataset	0.8339	0.8433	0.8581	0.8395
SMOTE_OVA	0.9193	0.9161	0.9182	0.9235
Bootstrap	0.9091	0.9091	0.8966	0.8706
HySM_BT40%*	0.9212	0.9369	0.9343	0.9158
HySM_BT50%*	0.9315	0.9336	0.9480	0.9315
HySM_BT60%*	0.9198	0.9145	0.9389	0.9127

Table 7: Comparison of G-means using hybrid data level approaches based on four different learning methods

Method	DL	SK	RF	GBT
Original dataset	0.7722	0.7841	0.8030	0.7796
SMOTE_OVA	0.8843	0.8799	0.8828	0.8901
Bootstrap	0.8704	0.8704	0.8538	0.8192
HySM_BT40%*	0.8869	0.9087	0.9051	0.8796
HySM_BT50%*	0.9012	0.9041	0.9242	0.9012
HySM_BT60%*	0.8850	0.8777	0.9115	0.8753

GBT). The instances obtained from combination sampling have the middle value in the data of a majority class and a minority class.

The output of this research shows that the hybrid data-level approaches with ensemble learning method (RF and SK) can help improve the efficiency of the model. Also, the accuracy rate is increased in a balanced sample of an elderly health condition dataset. The hybrid data level methods with ensemble learning provides the diversity of models which might reduce the bias of learners and decrease the skewed distribution for the imbalanced dataset classification.

In summary, all results clearly indicate that the hybrid data-level approaches based on OVA with different learning are applied to increase the efficiency of the data classification and to manage the imbalance of the elderly health condition data. Moreover, this technique still shows that it is an appropriate approach for the dataset with low imbalance ratio for multi-class.

The research also considers the procedural similarity hybrid data-level sampling techniques. From the previous studies [21] presents the combination SMOTE and OSS technique on EDM. SVM algorithm is used to predict the model and to improve accuracy of classification. The hybrid data-level solutions can help reduce the skew of each class and provide good effectiveness of classification in EDM. The differences from the previous studies are as follows; firstly, applying the hybrid data level approaches between SMOTE and bootstrap technique for elderly health condition dataset. In this process, there are resampling of instances approximately 40%- 60% of majority class to balance each class and to reduce the skew in the class

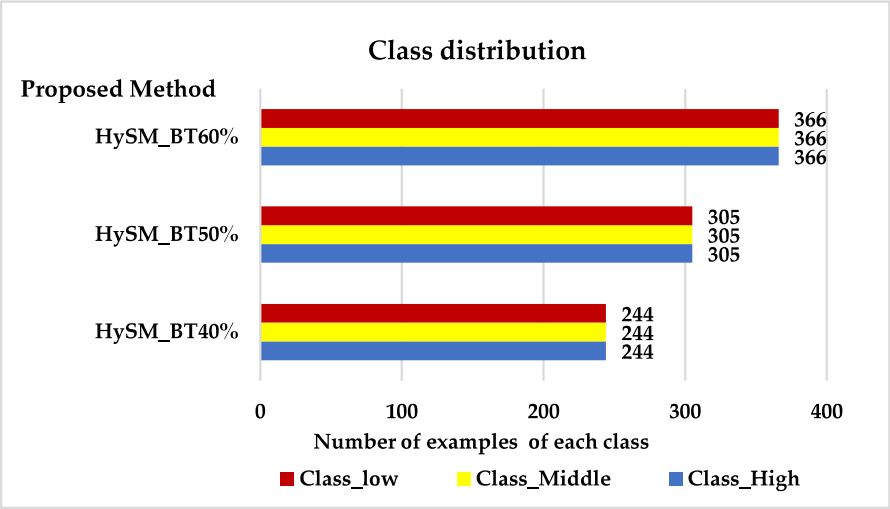

Figure 3: The distribution of instances per class using hybrid data level approach

distribution. Then, the data was applied to create the model for data classification. Secondly, in the process before evaluation of the classification model, the OVA strategy is applied to split the multi-class into binary classes. The validation splits up the example dataset into a training set (70%) and a test set (30%). Finally, four learning algorithms; deep learning, stacking algorithm, random forest, and gradient boosting tree are used in the evaluation process of the model classifier.

6 Conclusions

This study presents the hybrid method which the SMOTE and bootstrap sampling based on OVA technique with four different classifiers for the imbalanced multi-class management. The hybrid data-level approaches include HySM_BT40%, HySM_BT50%, and HySM_BT60%. The elderly's health condition data used in this research was obtained from Meaka Health Promotion Hospital in Thailand. It revealed that the efficiency of model classification has more accuracy. The HySM_BT50% method achieved the highest correctness when RF algorithm is used to classify the data. The best effective rate of classification with an accuracy of 90.11% (Sensitivity = 0.9011, Specificity = 0.9480, and G-means = 0.9242). The research results indicate that the combining sampling technique based on OVA strategy with DL, SK, RF, and GBT classifiers provides better classification accuracy rates than the single sampling approach because combination SMOTE and bootstrap sampling in the class imbalance is reduced. The advantage of the technique presented is to make the gained sample have the middle value between the data in the majority cluster and the data in the minority cluster, and helps increase the efficiency of the prediction model. This approach receives a prototype model for imbalanced multiclass handling. Additionally, the balance model is taken to create a prediction model to analyze the elderly health condition and a plan to promote elderly health care. Furthermore, the proposed approach might be a good choice for the dataset that has a large number of instances with low imbalance ratio. In the future, the feature selection method is applied in classifying the imbalanced data in order to select the feature having the importance and relating to each other. Moreover, the presented method might be applied to the other real-world problems.

Acknowledgments

This research project is supported by the Thailand science research and innovation fund and the University of Phayao (Grant No. FF66-RIM011). The Meaka Heath Promotion Hospital of Phayao in Thailand provided useful advice in data collection.

References

[1] A Agrawal, H. L. Viktor, and E. Paquet, "SCUT: Multi-Class Imbalanced Data Classification Using SMOTE and Cluster-Based Undersampling," *Proceedings of 7th International Joint Conference*, IC3K, Lisbon, Portugal, pp. 226-234, 12-14 November 2015.

[2] H. Ali, M. M. Salleh, K. Hussain, A. Ahmad, A. Ullah, A. Muhammad, and M. A. Khan, "Review on Data Preprocessing Methods for Class Imbalance Problem," *Int. J. Eng. Technol*., 8:390-397, 2019.

[3] H. Ali, M. M. Salleh, R. Saedudin, K. Hussain, and M. F. Mshtag, " Imbalance class prolems in data mining: A Review," IJEECS, 14:1560-1571, 2019..

[4] A. Csörgő Bentéjac, and G. Martínez-Muñoz, "A Comparative Analysis of Gradient Boosting Algorithms," *Artificial Intelligence Review*, 54:1937-1967, 2021.

[5] S. Desuky and S. Hussain, "An Improved Hybrid Approach for Handling Class Imbalance Problem," *Arab J Sci Eng*., 46:3853-3864, 2021.

[6] S. M. A. Elrahman and A. Abraham, "A Review of Class Imbalance Problem," *JNIC*, 1:332-340, 2013.

[7] E. Barrenechea Fernández, H. Bustince, and F. Herrera,

"An Overview of Ensemble Methods for Binary Classifiers in Multi-Class Problems: Experimental Study on One-vs-One and One-vs-All Schemes," *Pattern Recognition*, .44:1761-1776, 2011.

[8] S. García, J. Luengo, and F. Herrera, "Tutorial on Practical Tips of the Most Influential Data Preprocessing Algorithms in Data Mining," *Knowledge-Based Systems*, 98:1-29, 2016.

[9] A. Hanskunatai, "A New Hybrid Sampling Approach for Classification of Imbalanced Datasets," *Proceedings of 2018 3rd International Conference on Computer and Communication Systems (ICCCS)*, IEEE. Nagoya, Japan, pp. 67-71, 27-30 April 2018.

[10] R. Houari, A. Bounceur, A. K. Tari, and M. T. Kecha, "Handling Missing Data Problems with Sampling Methods," *Proceedings of International Conference on Advanced Networking Distributed Systems and Applications*, IEEE, Bejaia, Algeria, pp. 99-104, 17-19 June 2014.

[11] P. Kaur and A. Gosain, "Robust Hybrid Data-Level Sampling Approach to Handle Imbalanced Data During Classification," *Soft Computing,* 24:15715-15732, 2020.

[12] S. Kotsiantis, D. Kanellopoulos and P. Pintelas, "Handling Imbalanced Datasets: A Review," *GESTS International Transaction on Computer Science and Engineering*, 30:25-36, 2006.

[13] G. Kou, H. Chen, and M. A. Hefni, "Improved Hybrid Resampling and Ensemble Model for Imbalance Learning and Credit Evaluation," *JMSE*, 7:511-529, 2022.

[14] W. J. Lin, and J. J. Chen, "Class-Imbalanced Classifiers for High-Dimensional Data," *Briefings in Bioinformatics*, 14:13-26, 2013.

[15] C. Lorena, A. C. De Carvalho, and J. M. Gama, "A Review on the Combination of Binary Classifiers in Multiclass Problems," *Artificial Intelligence Review*, 30:19-37, 2009.

[16] B. Mahesh, "Machine Learning Algorithms-A Review, *Int. J. Sci. Res*., 9:81-386, 2020.

[17] R. Kora Mohammed, "A Comprehensive Review on Ensemble Deep Learning: Opportunities and Challenges," *Journal of King Saud University-Computer and Information Sciences*, 35:757-774, 2023.

[18] T. H. Oong and N. A. M. Isa, "One-Against-All Ensemble for Multiclass Pattern Classification," *Appl. Soft Comput.*, 12:1303-1308, 2012.

[19] K. Polat Ozdemir and A. Alhudhaif, "Classification of Imbalanced Hyperspectral Images using SMOTE-Based Deep Learning Methods," *Expert Syst. Appl.*, 178:114986, 2011.

[20] Y. Pandey, "Credit Card Fraud Detection using Deep Learning," *IJARCSSE*, 8:18-25, 2017.

[21] Y. Pristyanto, I. Pratama, and A. F. Nugraha, "Data Level Approach for Imbalanced Class Handling on Educational Data Mining Multiclass Classification," *Proceedings of 2018 International Conference on Information and Communications Technology,* IEEE*,* Yogyakarta, Indonesia, pp. 310-314, 06-07 March 2018.

[22] A. Pumsirirat and Y. Liu, "Credit Card Fraud Detection Using Deep Learning Based on Auto-Encoder and Restricted Boltzmann Machine," *Int J Adv Comput Sci Appl*, 9:18-25, 2018.

[23] E. Rendón, R. Alejo, C. Castorena, F. J. Isidro-Ortega, and E. E. Granda-Gutiérrez, "Data Sampling Methods to Deal with the Big Data Multi-Class Imbalance Problem," *Appl. Sci.*, 10:1-15, 2020.

[24] J. A. Sáez, B. Krawczyk, and M. Wozniak, "Analyzing the Oversampling of Different Classes and Types of Examples in Multi-Class Imbalanced Datasets," *Pattern Recognition*, 57:164–178, 2016.

[25] A. Shrestha and A. Mahmood, "Review of Deep Learning Algorithms and Architectures," *IEEE Access*, 7:53040-53065, 2019.

[26] Y. Sun, A. K. C. Wong and M. S. Kamel, "Classification of Imbalanced Data: A Review," *Int. J. Pattern Recognit. Artif.*, 23:687-719, 2009.

[27] C. Supromin and S. Choonhakhlai, "The Provision of Public Services in Municipalities in Thailand to Improve the Quality of Life of Elderly People," *Kasetsart Journal of Social Sciences*, 40:619-627, 2019.

[28] K. Upadhyay, Prabhjot Kakur, and S. Prasad, "A Review on Data Level Approaches to Address the Class Imbalance Problem," *Proceedings of International Conference on Recent Challenges in Engineering Science and Technology*, Andhra Pradesh, India, pp. 152-158, 9-10 April 2021.

[29] K. Upadhyay, P. Kaur, and D. K. Verma, "Evaluating the Performance of Data Level Methods using Keel Tool to Address Class Imbalance Problem," *Arab J Sci Eng*., pp. 1-14, 2022.

[30] S. Wan and H. Yang, "Comparison Among Methods of Ensemble Learning," *Proceedings of 2013 International Symposium on Biometrics and Security Technologies,* IEEE, Sichuan, China, pp. 286-290, 2-5 July 2013.

[31] Z. Xiang, Y. Su, J. Lan, D. Li, Y. Hu, and Z. Li, "An Improved SMOTE Algorithm Using Clustering," *Proceedings of 2020 Chinese Automation Congress (CAC) IEEE*, Shanghai, China, pp. 1986-1991, 06-08 November 2020.

[32] B. Xu, W. Wang, R.Yang, and Q. Han, "An Improved Unbalanced Data Classification Method Based on Hybrid Sampling Approach," *Proceedings of 2021 IEEE 4th International Conference on Big Data and Artificial Intelligence (BDAI),* IEEE, Qingdao, China, pp. 125-129, 2-4 July 2021.

[33] Z. Xu, D. Shen, T. Nie, Y. Kou, N. Yin, and X. Han, "A Cluster-Based Oversampling Algorithm Combining SMOTE and k-Means for Imbalanced Medical Data," *Information Sciences*, 572:574-589, 2021.

[34] Y. Zhang, J. Liu, and W. Shen, "A Review of Ensemble Learning Algorithms Used in Remote Sensing Applications," *Applied Sciences*, 12:1-20, 2022.

[35] Y. Zhao, Z. S. Y. Wong, and K. L. Tsui, "A Framework of Rebalancing Imbalanced Healthcare Data for Rare Events' Classification: A Case of Look-Alike Sound-Alike Mix-Up Incident Detection," *J. Healthc Eng.*, pp. 1-11, 2018.

Rattanawadee Panthong (photo not available) received the Ph.D. (Computer Science) from Kasetsart University, Thailand in 2021. She is currently an instruction at the School of Information and Communication Technology at the University of Phayao, Thailand. In addition, she is serving as an Assistant Dean in the School of Information and Communication Technology (2019–present). Her research interests are in machine leaning, data mining, data analytics and data warehouse. She can be contacted at email: rattanawadee.pa@up.ac.th.