Analysis of Oral Cancer Prediction with Pairwise Preprocessing Techniques using Hybrid Feature Selection and Ensemble Classification

Mumtazimah Mohamad, Nurul Athirah Rozlan, Fatihah Mohd

Abstract*: Class imbalance is one of main problem in data mining field that can prompt to misclassification. Data are said to be imbalanced if the classes instances are not appearing similarly. Despite the fact that the sample of the dominant class and their appropriate classification are vital to classifier, oral cancer is analyzed by depending on the minority class tests. Numerous classification learning algorithms have low prescient precision for the rare class. Additionally, majority of the classification algorithms concern on the classification of significant major sample while overlooking the minority class. Misclassification resulted to non-cancerous and the cancerous patients pay expansion time and cost. In this research study, an examination of imbalanced classification issue on oral cancer prediction will be thoroughly performed. This investigation utilizes crossover approach of SMOTE and Random Undersampling and mix of feature selection strategies. The proposed algorithm is expected to gives better class imbalance solution and better performance in classification of oral cancer prediction.

***Index Terms*: class imbalance, data preprocessing techniques, ensemble algorithm, feature selection.**

I. INTRODUCTION

Oral cancer usually caused by cancerous tissue growth with multiple factors from extrinsic to intrinsic that arise in a certain period of time. This disease incorporates malignant growths of mouth, cheek lining, gums, lips or palate, sinuses, and pharynx (throat), can be dangerous if not analyzed and treated early. The mortality rate related with this disease is high because of the malignant growth being found late in its improvement and it's not really recognizable in the beginning period. The present condition of dataset for these disease regularly high imbalance since majority of records dominated by critical patients.

Imbalanced class problem is vital from data mining's point of view. In medical data sets, data normally composed of minority and majority data and class. The imbalanced between both can cause classification problem where one or more class distribution is not balance in total with each other [6]. It regularly emerges in clinical data where high frequency among samples of cases that require frequent observation while case samples on the other hand may be at low frequency [2].

For an example, in the genuine medical records patients that have critical stage of cancer are more than the early stage cancer patients. Therefore, there will be more data for critical stage cancer patients rather than the early diagnosed. Subsequently, it is vital to determine which factors or feature that can prompt stage 4 oral cancer growth so as to get early diagnosis in diminishing the death rate of this disease. The class imbalance problem might be because of the restrictions to get information, for example, cost, protection and huge effort. However, in a lot of applications the minority class are the most important one. Direct classification might be biased towards the larger part of classes and results in poor execution in the minority class. Therefore, the cost of misclassifying the minority class correlated high rather with the cost of misclassifying the majority class. Therefore, the cost of classification error for minority class significantly high correlated with the cost of classification error for majority class.

All learning algorithms assumed that there would be a similar size in data set as a consequence that the learning algorithms cannot model the minority class. It is because in many time it neglects the rare class. Therefore, the results will significantly represent the majority class since minority class contribute less [7]. In data training, this imbalance class will give some effects to the execution for mostly selected cases such as higher error rate and the low percentage of accuracy [8], [9]. To improve this problem, a resampling technique imbalanced data set prior to feature selection step taking place is proposed. Hybrid approach which is a combination of random undersampling and SMOTE were used in our model to resolve the imbalance problem.

Most data contain more information than is needed to build the model, or the less important information. Feature selection technique is one of the good technique that can help us to create an accurate predictive model by removing the unneeded, redundant and irrelevant attributes that do not contribute to the accuracy of the model or may lead to the lower accuracy of the predictive model. In this study, this method was applied to select the optimum feature in order to enhance model accuracy.

II. RELATED WORK

Classification is one of the data mining technique used to target classifications or classes. The aim of classification is to predict accurately the target class for each case in the data. The classification for head and neck mostly considers machine learning and data mining algorithm such in [9] that predict the oral cancer patients by employing clustering K-Mean and neural network with histopathological data

Revised Manuscript Received on February 11, 2019.

Mumtazimah Mohamad, Faculty Informatics and Computing, Universiti Sultan Zainal Abidin, Besut, Malaysia.

Nurul Athirah Rozlan, Faculty Informatics and Computing, Universiti Sultan Zainal Abidin, Besut, Malaysia.

Fatihah Mohd, School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu, K. Terengganu, Malaysia.

set. The model achieve accuracy of more than 90 % for the diagnosed results. In [10] shows distinct three unique of decision tree the predicting of survivability is higher and accuracy is increased. In many case, classification strategies may unsuccessful or unreliable when experimented to an imbalanced dataset [7].

Methods that can used to solve imbalanced data are categorized in three approaches [11]. The first approach is by algorithm level where this method focuses on the learning of minority class that map the classifier algorithm to improve performance the present of minority class [1]. The second approach is by data level, where this technique aims to solve problems with the distribution of a data set by using sampling methods [12]. Oversampling and undersampling are among basic sampling where oversampling copies minority class cases in random, while undersampling cast aside the majority class cases randomly to enhance the class distribution. Oversampling may lead to overfitting as duplicate the minority cases while undersampling may discard potential important majority cases. Third approach is by using cost-sensitive learning framework. This technique make use between algorithm and data approaches.

Apart to the approaches, the ensemble learning emerges to be employed for imbalance data. Ensemble methods are collection of algorithm that join a few machine learning techniques into one predictive model. There are several techniques that commonly used as an ensemble technique such bagging, boosting, stacking and etc. In a few research, this strategy using several techniques such as clustering algorithm and genetic algorithm and might resulted better performance than bagging and boosting [13].

From the above categorization, the three kind of strategies have been utilized and improved and still under emerging research activities. Some research proposed an enhanced feature selection technique with machine learning such as SVM and RBF [14]. Solution for variety sampling and resampling also commonly used [15] for example using two non-random instances of informed undersampling that demonstrated to give great outcomes are EasyEnsemble and BalanceCascade algorithms [16]. On the other side, hybrid approach also be used as an integration of oversampling and undersampling by remove a portion of the example previously or after to resampling in order to handle overfitting [17]. Correlation-based feature selection and the wrapper algorithms also being utilized Bayesian Networks, Artificial Neural Networks, SVM, DTs and Random Forests improved a bit the performance [18]. A hybrid model of Relief Genetic Algorithm with ANFIS delivered a higher classification accuracy more than 90% [19]. All of these research mostly being studied based on data mining Knowledge Data Discovery (KDD) as the framework consist of data cleaning, data transformation, data mining, pattern evaluation and knowledge presentation.

III. METHODOLOGY

To accomplish our research objectives, the research will be divided into five main phases. The flow of the proposed methods were shown in Fig. 1 and will be discussed in next subsection.

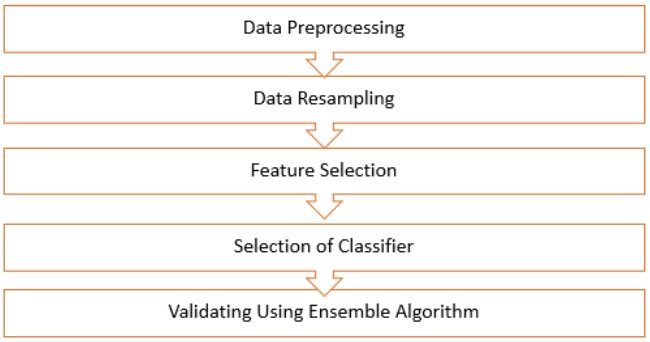

Fig. 1: Process flow

A. Data Preprocessing and Resampling

Data normalization is used in the data preprocessing stage where the original data will be normalized and missing data will be treated. SMOTE-RUS resampling were implemented in the second phase of experiment. Synthetic Minority Oversampling Technique (SMOTE) copying insignificant data and oversample it by making synthetic samples in the feature space shaped by the occurrence and its *K*-nearest neighbors, which adequately dodge the overfitting issue [22]. For a given two samples, x_1 and x_2 from random minority sample set, where each sample has n attributes. For x_1 and x_2, calculate the difference of i_{th} attribute; which is, $diff_i = x_{2i} - x_{1i}$. Then, obtain the i_{th} attribute value of the new target sample according to

$$x_{12i} = \text{r and } [0, 1] * diff_i \quad (1)$$

where r and [0, 1] an arbitrary value in the range of 0 and 1. So, the final synthetic sample of x_1 and x_2 is

$$x_{2i} = \text{r and } [0, 1] * \quad diff \quad (2)$$

where $diff = (diff_1, diff_2, \ldots . diff_n)$ according to the sampling rate in certain execution times and repeated. The synthetic samples and the original samples is used as final minority sample set. Then, random undersampling techniques and SMOTE were combined and implemented to individually increment or reduce the size of the classes to accomplish the ideal proportions [25]. SMOTE utilized five closest neighbors, which was also compatible with results as in [4]. Given a dataset S, with minority class S_P and larger part class S_N, the system can be depicted as follows:

i. The modified size, *newMajSize*, of the larger part class, is defined by a random number generated between 2 and |S|−2 (both inclusive). Appropriately, the modified measure, *newMinSize*, of the minority class becomes |S| − *newMajSize*.
ii. If *newMajSize* < $|S_N|$, the larger part class S'_N is made by RUS the original S_N so that the final measure $|S'_N|$ = *newMajSize*. Thus, the new minority class S'_P is acquired from S_P utilizing SMOTE to create *newMinSize* − $|S_P|$ artificial instances.
iii. Otherwise, S'_P is the class created by RUS S_P. On the other hand, S'_N is the class that includes artificial samples generated using SMOTE on S_N. Thus, finally, $|S'_P|$ = newMinSize and $|S'_N|$ = *newMajSize*.

B. Hybrid Feature Selection Approach

The aim of feature selection technique is to locate most the most important attributes as this will contribute to the performance of the classification. It is critical to apply feature selection since class imbalance problem is ordinarily joined by issue of high data dimensionality. In this project, feature selection implemented in WEKA with 10-fold cross validation. The function utilized for attribute assessment will be discussed in next section.

Correlation Attribute Evaluator with Ranker. This algorithm valuates an attribute by estimating the connection of all attribute with the class. This nominal characterization is carry out based on value as an indicator where general relationship for nominal attribute is accomplished by a weighted average. This algorithm utilized with ranker algorithm, so as to rank its traits by using individual assessments.

Correlation Forward Selection Subset Evaluator with Linear Forward Selection. This algorithm evaluates the value of a subset of attributes by deliberating the individual predictive capability of each element unto the level of iteration between them. This forward selection considers a limited number of k characteristics. Each set chooses a fixed number k of traits, while k is expanded in each progression when fixed-width is chosen. The search is used to utilize for initial ordering to select the top k attributes, or to do a ranking base on similar value but can be used later.

C. Ensemble Classification

The fundamental point of created the ensemble classifiers is to improve performance of classifier [21]. Boosting strategy is an ensemble learning algorithm that can tackle the issue of class imbalance and to improve the performance of weak classifier [22]. Four classifiers were selected to the classification which are SVM, Naïve Bayes, Logistic Regression and KNN. An ensemble algorithm which is AdaBoost implemented to validate the analysis performance for each of the classifier when experimented with different resampling and feature selection techniques. The AdaBoost ensemble algorithm were utilized to make a relative study with the four single classifier.

The fundamental rule of boosting is to fit a grouping of weak learner, for example, models that are just somewhat superior to random guessing. In this algorithm, more weight is given to example that were misclassified by prior rounds. The prediction are then consolidated through a weighted majority vote (classification) to create the last forecast. AdaBoost depends on the rule of creating numerous predictors and weighted election among the individual feeble classifier.

AdaBoost continues in a successive strides with equivalent probabilities of each {D(X_i)}, i = 1,…, n in the learning dataset at the absolute starting point. At each progression, the new learning dataset is chosen by examining from the first learning set utilizing probabilities {D(X_i)}, i = 1,…, n with substitution. After the classifier dependent on this resampled learning set is built, the case {D(X_i)}, (i = 1,…, n) are refreshed relying upon the misclassifications up to the present step. That is, the focuses being misclassified in preceding step will be allocated heavier sampling probabilities and have huge opportunity to be chosen into the learning set for subsequent step.

Algorithm AdaBoost

Input: sequence of m examples $\langle(x_1, y_1), \ldots, (x_m, y_m)\rangle$ with clasess

$y_i \in Y = \{1, \ldots, k\}$

weak learning algorithm **WeakLearn**

integer T specifying number of iterations

Initialize $D_1(i) = 1/m$ for all i.

Do for $t = 1, 2, \ldots, T$

1. Call **WeakLearn**, providing it with the distribution D_t
2. Get back a hypothesis $h_t : X \rightarrow Y$
3. Calculate the error of h_t: $\epsilon_t = \sum_{i:h_t(x_i)\neq y_i} D_t(i)$. If $\epsilon_t > \frac{1}{2}$, then set $T = t - 1$ and abort loop.
4. Set $\frac{\beta_t = \epsilon_t}{1-\epsilon_t}$.
5. Update distribution D_t: $D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} \beta_t & \text{if } h_t(x_i) = y_i \\ 1 & \text{otherwise} \end{cases}$ where Z_t is a normalization constant (chosen so that D_{t+1} will be a distribution).

Output the final hypothesis:

$h_{fin}(x) = \arg max_{y\in Y} \sum_{t:h_t(x)=y} log \frac{1}{\beta_t}$

Fig. 2: The algorithm AdaBoost

The boosting calculation carry out as information on a preparation set of m models $S = \langle(x_1, y_1), \ldots, (x_m, y_m)\rangle$ where x_1 is an occurrence depicted from some space X and drawn to certain way such as in vector, and $y_i \in Y$ is the class related with x_i. In this study, the arrangement of conceivable names Y is of limited cardinality k is generally expected.

In addition, the boosting algorithms approaches another learning algorithm, called the weak algorithm (weakLearn). In this study, SVM, Naïve Bayes, Logistic Regression and KNN classifier used as a weakLearn to the AdaBoost. The boosting algorithms calls weakLearn repetitively in a progression of rounds. On round t, the supporter gives weakLearn a dissemination D_t over the preparation set S. Accordingly, weakLearn registers a classifier or theory $h_t : X \rightarrow Y$ which ought to misclassify an important vector of the preparation models, in respect to D_t. That is, the weak learner's goal will likely discover a theory h_t which reduce the error $\epsilon_t = PR_i \sim D_t [h_t (x_i) \neq y_i]$. Note that this error is estimated regarding the dissemination D_t that was given to the weak learner. This procedure proceeds for T rounds, and, finally, the booster integrate the weak hypothesis $h_1, \ldots\ldots, h_T$ into a single last hypothesis h_{fin}.

IV. EXPERIMENTAL SETUP

The resource data for this study were obtained from a record perspective on oral disease patients from Otorhinolaryngology Clinic at Hospital Universiti Sains Malaysia (HUSM) in Kelantan. The data set consist of 27 attributes. The variable in the data set were sex, ethnic, age gathering, oral malignancy district and high-chance propensities. Age was partitioned into five categories. The oral malignant growth districts incorporated into this research

were buccal, floor of mouth, lip, mandible, maxilla, Oropharynx, sense of taste, salivary organ, tongue and other unspecific parts. The model of attributes found in this dataset are recorded in Table 1.

Table 1: Oral cancer dataset consist of 27 variables

No.	Variable	Type
1.	Case Id	Text
2.	Age	Numeric
3.	Gender	Nominal
4.	Ethnicity	Polynomial
5.	Smoking	Nominal
6.	Quid chewing habit	Nominal
7.	Alcohol	Nominal
8.	Difficulty in chewing	Nominal
9.	Ulceration	Nominal
10.	Neck lump	Nominal
11.	Loss of appetite	Nominal
12.	Loss of weight	Nominal
13.	Hoarseness of voice	Nominal
14.	Bleeding	Nominal
15.	Burning	Nominal
16.	Painful	Nominal
17.	Swelling	Nominal
18.	Numbness	Nominal
19.	Site	Polynomial
20.	Size	Polynomial
21.	Lymp node	Nominal
22.	Histopatological type	Polynomial
23.	Differentiation of SCC	Polynomial
24.	Primary Tumou	Polynomial
25.	Regional lymph nodes	Polynomial
26.	Distant Metastasis	Polynomial
27.	Stage	Polynomial

V. RESULTS AND DISCUSSION

A. Resampling Approach

The hybrid approach of SMOTE and random undersampling duplicated minority data by 5 times and haphazardly discard rows from majority class to achieve similar number of instances for each class. The result of balanced class distribution for oral cancer data set were shown in Table 2.

Table 2: Balanced Class Distribution for OC utilizing hybrid approach by SMOTE and Random undersampling (SMOTE-RUS)

Class Name	Without Resampling		After SMOTE-RUS	
	# Instances	% (Instances)	# Instances	% (Instances)
One	4	4.89	20	25.00
Two	10	12.20	20	25.00
Three	28	34.14	20	25.00
Four	40	48.78	20	25.00
Total	82		80	

The result shows that the SMOTE-RUS resampling balance the class equally by 20 instances for each class. SMOTE oversampling the minority class by synthetically generating data points. In this study, we use 5 nearest neighbors for the amount of oversampling for the minority sample that was chosen randomly. Then, the random undersampling arbitrarily removed the line of the majority class until each class had equally number of sample as in the majority class.

B. Optimum Feature Selected

The algorithm started with the origin oral cancer data set which contain 27 features and 82 instances. The feature selection are carried out in WEKA with 10-fold cross validation. Correlation Ranking Filter resulted to ranking 27 features which are 23, 19, 24, 17, 25, 16, 2, 22, 3, 9, 20, 5, 21, 1, 18, 10, 13, 11, 4, 14, 12, 8, 15, 7, 26. In this process, 3 features with ranking rate 0 were removed which 7, 15, 26. Then, with the balance of 24 features CFS Subset Eval remove irrelevant 18 features namely 23, 19, 24, 25, 16, 2, 3, 9, 20, 5, 1, 18, 10, 13, 11, 4, 14, 12, 8, 6.

Table 3 demonstrates the hybrid feature selection implemented in this research study. The algorithm began with no feature selection (FS0), Correlation Attribute Evaluator with Ranker (FS1). At that point, the hybrid technique occurred when FS1 integrated with CFS Subset Evaluator with Linear Forward Selection (FS2). Based on Table 3, the aim for the feature selection techniques was to select the most relevant attributes that can contribute to the output class. We actualized FS1 to compute the connection between each attribute and the output variable and select only those attributes that have a moderate-to-high positive or negative relationship (near 1 or 1) and evacuate those traits with a low connection (zero esteem or near zero). The outcome demonstrates that attribute 23 has the most noteworthy relationship with the output class and 3 attribute with zero connection with the output class were evacuated.

Table 3: Attributes selection with feature selection technique

FS	Method	Selected Attributes
FS0	No selected feature	27 attributes
FS1	Correlation Attribute Eval Ranker	Ranked Attribute: 23,19,24,17,25,16,2,22,3,9,20,5,21,1,18,10,13,11,4,14,12,8,6,15,7,26 (27 attributes) Remove ranting value : 0 (attribute 7,15,26)
FS2	cfsSubSetEval Linear Forward Selection	Remove irrelevant attributes: 18 attributes 23,19,24,25,16,2,3,9,20,5,1,18,10,13,11,4,14,12,8,6 Optimum features: 5 attributes 1,15,17,21,22

Then, with the attributes left by the FS1 we executed CFS Subset Eval with Linear Forward Selection to the left attributes in the FS1 to look through the most relevant features resulting in optimum features chose with 5 attributes which is 1, 15, 17, 21, 22. This combination of two feature selections resulting in a hybrid process.

C. Accuracy Classification Performance

The performance measure is considered to assess the productivity of the SMOTE-RUS AND FS techniques. Basically the precision is the most used measurement to compute the performance of classifiers [2]. The performance of this methods is assessed by confusion matrix functional evaluation measure.

The confusion matrix is a helpful method to prove the result of a classifier since every one of the measurements used to assess classifiers can be figured from it [22]. Confusion matrix include the genuine and anticipated characterization of a classifier. Depending on the data, the execution of a classifier is assessed by confusion matrix as in Table 4. The assessment are gathered by the Classification Accuracy (%) = (TP + TN)/(TP + FP + FN + TN). In this paper, the minority class is positive and the dominant part class is negative.

Four classifier were used in the classification which is SVM, Naïve Bayes, Logistic Regression and KNN. For KNN and SVM all string or class feature were standardized and mapped to numerical qualities where essential. The information for the esteem mapping was taken from the dictionaries contained in a Predictive Model Markup Language records (PMML) and combined with experimental environment. [25]. Naive Bayes classifier using estimator class. Logistic regression predicts probability and models the probability of the default class.

Table 4: Evaluation metrics for imbalance learning

Performance Rate	Formulation
True positive rate	$TP_{rate} = \frac{TP}{TP+FN}$ is the percentage of positive instances classified correctly.
False positive rate	$FP_{rate} = \frac{FP}{FP+TN}$ is the percentage of negative instances misclassified.
False negative rate	$FN_{rate} = \frac{FN}{TP+FN}$ is the percentage of positive instances misclassified.
True negative rate	$TN_{rate} = \frac{TN}{FP+TN}$ is the percentage of negative instances classified correctly.

In this research, these four diverse machine learning algorithms were utilized to group the oral cancer growth data set with three feature selection techniques and optimum features were chosen by the hybrid approach. Table 5 shows results of the percentage of accuracy for selected different feature selection without resampling. The approach resulted to reduction of accuracy from FS1 to hybrid approach FS2. Naïve Bayes show the highest classification accuracy with FS1 with 78.05%. Next score was obtained by KNN with 74.39%. SVM has the worst accuracy with remaining 48.78% for all three feature selection. Based on the result, the accuracy for hybrid feature selection are mainly not perform the best without resampling technique.

Table 5: Accuracy result for selected feature selection on OC data set without resampling

Classification Accuracy Without Resampling (%)				
Algorithm		FS0	FS1	FS2
Support Vector Machine	Accuracy	48.78	48.78	48.78
	Incorrectly classified	51.22	51.22	51.22
Naïve Bayes	Accuracy	78.05	78.05	70.73
	Incorrectly classified	21.95	21.95	29.27
Logistic Regression	Accuracy	63.41	63.41	53.66
	Incorrectly classified	36.58	36.59	46.34
k-Nearest Neighbours	Accuracy	74.39	74.39	70.73
	Incorrectly classified	25.61	25.61	29.57

Meanwhile, in Table 6 shows the performance of Selected FS with SMOTE- RUS. The approach improved the accuracy performance for all classification algorithms with accuracy higher than 95%. The best performing feature selection techniques was by hybrid feature selection FS2 with SVM algorithm at 98.75% accuracy. The lowest accuracy was performed by Logistic Regression in FS1 with 96.25% and Naïve Bayes in FS2 also with the same accuracy percentage. Most of the classifiers with all feature selection techniques performed well when combined with SMOTE-RUS resampling. Based on the results obtained, there are large differences on the accuracy without resampling and resampling method. All the classification algorithm shows an accuracy improvement when implemented with resampling method. For the classification with SVM, the accuracy increased almost 50% when implemented with SMOTE-RUS rather than without resampling.

Table 6: Accuracy result for selected feature selection on OC data set with SMOTE-RUS

Classification Accuracy With SMOTE-RUS (%)				
Algorithm		FS0	FS1	FS2
Support Vector Machine	Accuracy	97.50	97.50	98.75
	Error	2.50	2.50	1.25
Naïve Bayes	Accuracy	97.50	97.50	96.25
	Error	2.50	2.50	3.75
Logistic Regression	Accuracy	96.25	96.25	97.50
	Error	3.75	3.75	2.50
k-Nearest Neighbours	Accuracy	97.50	97.50	97.50
	Error	2.50	2.50	2.50

Based on Table 6, the hybrid FS approach used which are represented by FS2 shown that classification accuracy improved from 97.50% to 98.75% for SVM, 96.25% to 97.50% for Logistic Regression and remain the same for KNN which 97.50% but the accuracy for Naïve Bayes decreased about 1.25% compared to FS1. SVM shows the highest classification accuracy performance with 98.75% for FS2. While in Table 5, SVM has the worst accuracy when performed with hybrid FS2 with accuracy 48.78%. These show the resampling method could improve the classification power for SVM algorithm.

D. Validating Results using Ensemble Method

Experimental results from Table 7 present that data feature selection technique only able to enhance the accuracy of classification for Logistic Regression from 63.41% to 68.29% when implemented with AdaBoost without SMOTE-RUS resampling technique. Meanwhile, the accuracy for Naïve Bayes and KNN decreased from 71.95% to 69.51% and 74.39% to 70.73% respectively when experimented without resampling along with hybrid feature selection. The accuracy for SVM shows no improvement with remain the same at 48.78%.

Based on Table 5 and Table 7, without resampling method Logistic Regression shows an accuracy improvement from 53.66% to 68.29% when using AdaBoost but Naïve Bayes accuracy decreased by 1.7% from 70.73% to 69.51% while the accuracy for SVM and KNN not showing any improvement.

From Table 7 and 8, with AdaBoost algorithm SVM show the significant impact of SMOTE-RUS resampling technique where the accuracy increase almost by 50% from 48.78% to 96.25%. From these results, we obtained a high improvement of accuracy which all classification algorithms have an accuracy higher than 95% when data set implemented with SMOTE-RUS resampling technique with AdaBoost ensemble algorithm rather than without resampling technique. Meanwhile, the hybrid feature selection technique, FS2 able to improve the accuracy for SVM and Logistic Regression from 95.0% to 96.25% and 96.25% to 97.50% respectively. The highest accuracy obtained by SVM with 98.75% using SMOTE-RUS resampling technique with hybrid feature selection FS2 in Table 6.

Table 7: Accuracy result without resampling for selected feature selection on OC data set using AdaBoost

Classification Accuracy without Resampling using AdaBoost.M1 (%)				
Algorithm		FS0	FS1	FS2
Support Vector Machine	Accuracy	48.78	48.78	48.78
	Error	51.22	51.22	51.22
Naïve Bayes	Accuracy	71.95	71.95	69.51
	Error	28.05	28.05	30.49
Logistic Regression	Accuracy	63.41	63.41	68.29
	Error	36.58	36.58	31.70
k-Nearest Neighbours	Accuracy	74.39	74.39	70.73
	Error	25.61	25.61	29.27

Table 8: Accuracy result with resampling for selected feature selection on OC data set using AdaBoost

Classification Accuracy With SMOTE-RUS using AdaBoost.M1 (%)				
Algorithm		FS0	FS1	FS2
Support Vector Machine	Accuracy	96.25	95.00	96.25
	Error	3.75	5.00	3.75
Naïve Bayes	Accuracy	97.50	97.50	96.25
	Error	2.50	2.50	3.75
Logistic Regression	Accuracy	96.25	96.25	97.5
	Error	3.75	3.75	2.50
k-Nearest Neighbours	Accuracy	97.5	97.50	97.5
	Error	2.50	2.50	2.50

These result proved that the SMOTE-RUS resampling solved the imbalance class problem with the accuracy improvement for all classification algorithm either when implemented with or without AdaBoost. The results also show that the hybrid feature selection FS2 only show an improvement when data set experimented with resampling technique. In the classification phase, the ensemble methods are utilized to enhance the accuracy but in this study, ensemble methods used to validate the impact of resampling technique and hybrid feature selection from the different classifier.

VI. CONCLUSION

In this paper, a pairwise integration of data preprocessing and hybrid feature selection for taking care of imbalance data is proposed. The utilization of resampling strategies for imbalanced data set successfully accomplishing higher outcomes in term of accuracy rather than without resampling. The validation result using AdaBoost.M1 also shows that all classification algorithms have an improved accuracy when implemented with SMOTE-RUS resampling technique rather than boosting algorithm. Our future work will include cost-sensitivity as a part of solution for class imbalance.

ACKNOWLEDGMENT

This work was supported by the Fundamental Research Grant Scheme (FRGS/1/2015/ICT02/UNISZA/02/1) by Malaysia Ministry of Higher Education and the Center of Research and Innovation Management of Universiti Sultan Zainal Abidin, Terengganu, Malaysia.

REFERENCES

1. Z. A. Bakar, F. Mohd, N. Maizura M. Noor, and Z. A. Rajion, "Demographic profile of oral cancer patients in East Coast of Peninsular Malaysia," International Medical Journal, 20(3), 2013, pp. 362-364.
2. A. Wosiak, and S. Karbowiak, "Preprocessing compensation techniques for improved classification of imbalanced medical datasets," IEEE Federated Conference on Computer Science and Information Systems, 2017, pp. 203-211.
3. S. Wilk, J. Stefanowski, S. Wojciechowski, K. J. Farion, and W. Michalowski, "Application of preprocessing methods to imbalanced clinical data: An experimental study," Conference of Information Technologies in Biomedicine, 2016, pp. 503-515.
4. F. Mohd, Z. A. Bakar, N. M. M. Noor, Z. A. Rajion, and N. Saddki, "A hybrid selection method based on HCELFS and SVM for the diagnosis of oral cancer staging," Lecture Notes in Electrical Engineering, 315, 2015, pp. 821-831.
5. M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(4), 2012, pp. 463-484.
6. G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," Expert Systems with Applications, 73, 2017, pp. 220-239.
7. B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," Progress in Artificial Intelligence, 5(4), 2016, pp. 221-232.
8. V. García, J. S. Sánchez, and R. A. Mollineda, "On the effectiveness of preprocessing methods when dealing with different levels of class imbalance," Knowledge-Based Systems, 25(1), 2012, pp. 13-21.
9. W. T. Tseng, W. F. Chiang, S. Y. Liu, J. Roan, and C. N. Lin, "The application of data mining

techniques to oral cancer prognosis," Journal of Medical Systems, 39(5), 2015, pp. 1-7.
10. N. Sharma, and H. Om, "Data mining models for predicting oral cancer survivability," Network Modeling Analysis in Health Informatics and Bioinformatics, 2(4), 2013, pp. 285-295.
11. S. Wang, and X. Yao, "Relationships between diversity of classification ensembles and single-class performance measures," IEEE Transactions on Knowledge and Data Engineering, 25(1), 2013, pp. 206-219.
12. C. Zhang, J. Bi, and P. Soda, "Feature selection and resampling in class imbalance learning: Which comes first? An empirical study in the biological domain," IEEE International Conference on Bioinformatics and Biomedicine, 2017, pp. 933-938.
13. N. A. Abolkarlou, A. A. Niknafs, and M. K. Ebrahimpour, "Ensemble imbalance classification: Using data preprocessing, clustering algorithm and genetic algorithm," 4th International Conference on Computer and Knowledge Engineering, 2014, pp. 171-176.
14. P. Jaganathan, N. Rajkumar, and R. Kuppuchamy. "A comparative study of improved F-score with support vector machine and RBF network for breast cancer classification," International Journal of Machine Learning and Computing, 2(6), 2012, pp. 741-745.
15. H. He, and E. A. Garcia, "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, 21(9), 2009, pp. 1263-1284.
16. R. F. A. B. D. Morais, P. B. C. Miranda, and R. M. A. Silva, "A meta-learning method to select under-sampling algorithms for imbalanced data sets," 5th Brazilian Conference on Intelligent Systems, Recife, 2016, pp. 385-390.
17. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, 16, 2002, pp. 321-357.
18. K. P. Exarchos, Y. Goletsis, and D. I. Fotiadis, "Multiparametric decision support system for the prediction of oral cancer reoccurrence," IEEE Transactions on Information Technology in Biomedicine, 16(6), 2012, pp. 1127-1134.
19. S. W. Chang, S. A. Kareem, A. F. Merican, and R. Zain, "Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods," BMC Bioinformatics, 14(1), 2013, pp. 1-15.
20. S. Ertekin, J. Huang, and C. L. Giles, "Active learning for class imbalance problem," Special Interest Group on Information Retrieval Conference, 2007, pp. 823-824.
21. R. Longadge, and S. Dongre, "Class imbalance problem in data mining review," International Journal of Computer Science and Network, 2(1), 2013, pp. 1-6.
22. E. Olivetti, S. Greiner, and P. Avesani, "Statistical independence for the evaluation of classifier-based diagnosis," Brain Informatics, 2(1), 2015, pp. 13-19.
23. C. Gong, and L. Gu, "A novel SMOTE-based classification approach to online data imbalance problem," Mathematical Problems in Engineering, 2016, 2016, pp. 1-14.
24. V. García, J. S. Sánchez, and R. A. Mollineda, "On the effectiveness of preprocessing methods when dealing with different levels of class imbalance," Knowledge-Based Systems, 25(1), 2012, pp. 13-21.
25. D. Morent, K. Stathatos, W. C. Lin, and M. Berthold, "Comprehensive PMML preprocessing in KNIME," Workshop on Predictive Markup Language Modeling, 2011, pp. 28-31.