Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2182 – 2190
Received 05 March 2021; Accepted 01 April 2021.

Tomek link Undersampling with Stacked Ensemble classifier for Imbalanced data classification

Kamaladevi M[1], **Venkataraman V**[2*],**Sekar K R**[3]

[1]SASTRA Deemed University Srinivasa Ramanujan Centre Kumbakonam TamilNadu,India
[2,3]SASTRA Deemed University Thanjavur TamilNadu India
*mathvvr@maths.sastra.edu

ABSTRACT

Imbalanced data is referred as unequal distribution of data between classes as a result of classifier. Majority class have higher percentage of data than minority class .Predicting the classifier accuracy is mostly biased toward the majority class, this leads to class imbalance problem. Misprediction of minority class has great loss in financial sector and life threatening problem in medical field. To solve the issue imbalanced dataset should be balanced by means of oversampling the minority class data or undersampling the majority class data. Oversampling can cause over fitting of data. Among many undersampling algorithm Tomek link under sampling algorithm remove the noisy and borderline samples from majority class without losing important information of data. Tomek link undersampling algorithm can be applied to 4 benchmark Imbalanced dataset from UCI repository such as breast-W, breast cancer, hepatitis, heart diseases and so on. Stacking Ensemble classifier combine the output of individual classifier and fed to next level meta classifier to predict the result. First level prediction has been done using individual classifier such as NaiveBayes, Logistic Regression etc the output of first level fed to Meta classifier which predict the final class label. Performance measures such accuracy, precision, recall and auc_roc score are measured and compared with state of art classifier such as Support vector Machine, Decision tree, Naïve Bayes Classifier, K-Nearest Neighbour .Tomek link undersampling data applied to stacking classifier predict accuracy of 0.94 and roc_auc score of 0.97 which outperform the individual classifier performance measure

Keywords
Tomek link Undersampling, UCI , Support Vector Machine, Logistic Regression, Naïve Bayes ,K-Nearest Neighbour

Introduction

Imbalanced data surprisingly a common problem in machine learning algorithm especially in classification. Classifier result in uneven distribution of data in two classes. Majority class contains more data than Minority class. Classification Algorithm performance are measured using their accuracy. Most of the performances measures for classifier are ruled out for imbalanced classification. Accuracy measured is turn towards the majority class concealing the minority class accuracy. It creates a great impact in some data such as finance and medical. Imbalanced data set are available in following domain such as Diseases Diagnosis, Fraud detection, spam filtering etc., Most of the classifier are calculating overall accuracy instead of measuring the individual class accuracy. Among various application, diseases diagnosis is the common problem in all over the world. Early detection of diseases saves human lives. Predicting diseases through machine learning algorithm give accurate result than human prediction. Many of the physicians are make use of this disease detection method as a evidence to their treatment.

For example in disease diagnosis data, classifier predict that people having disease or not in early stages. Most of the people don't have diseases consider as majority class and few have suffered from diseases that can be taken as minority class .Accuracy for minority class is not properly evaluated, which result in wrong prediction .People who have suffered from diseases can identify as healthy people and they can not get proper treatment. In order to overcome the drawback ,Imbalanced problems are handled by three ways i)Data approach ii)Algorithmic approach iii)Hybrid approach.

Data approach make the imbalanced data class distribution as a balanced one. Balancing the data done by two ways

- OverSampling
- Undersampling

Oversampling means generate new samples for minority class which will almost equal to number of samples of majority class. .Adding new sample will result in overfitting. Random Oversampling and SMOTE are famous oversampling technique.

Undersampling technique get rid of some of the samples in majority class in order to balance the class distribution .Imbalanced data classes ratio are normally 100:20,80:10.Undersampling reduce the skewed ratio to

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2182 – 2190
Received 05 March 2021; Accepted 01 April 2021.

40:20,20:10.Randomly remove the samples from majority class results in losing important information of data.This limitation should be overcome by selecting samples using some mathematical techniques will help for induction process. Heuristic decision on sample improve the data standard .
Neighbourhood undersampling select instances from majority class based on distances between two samples[]. Undersampling with ensemble algorithm predict classes in more accurate manner.

Literature Review

Imbalanced classification problem can be improved by Ensemble classifier. It performance are measured using ROC_AUC curve. Data are preprocessed using sampling technique before applying to Ensemble classifier. SMOTE algorithm used for oversampling the minority class will reduce the imbalance problem[1]. B-stacking approach used for credit scoring model. Classifier pool are generated and finally fusion is used to select the classifier from ensemble and pool to give better classifier[2]

In Ensemble Classification weighting mechanism are added to minority class instance. To improve the diversity in classifier DES algorithm are used .Meta learning Framework of Ensemble improve accuracy [3]. In Predicting heart diseases for small dataset Naïve bayes algorithm give better result. When the dataset is big Decision tree give good result. Percentage of heart disease are also able to calculated using machine learning algorithm[5].Doctors need some support system to predict the heart attack using machine learning algorithm. Deep learning algorithm also provide new door for predicting heart diseases[4].Different machine learning algorithm such as support vector machine,C4.5 decision tree and K-nearest neighbor are applied to dataset and finally select best algorithm with more accuracy. Optimization of base classifier using PSO algorithm make an Ensemble classifier to improve the prediction. Integrating multiple classification algorithm and through average voting design a best Ensemble. Adaboost algorithm distribute weight evenly to all training data and run in base learner and misclassified data weights are adjusted for next training iteration. Neural network are train the data by increasing the number of hidden layer. Back propagation neural network are used to train data and applied to adaboost algorithm to give high accuracy[6].

Support vector Machine classification algorithm is modified as Boosted SVM. Modification can be done in two ways distance based weight updated for boosting algorithm and classifier removal based on sign .Comparing modified boosted algorithm with other classifier shows better performance[7]. Undersampling technique are available for rebalance the data. Clustering based undersampling make cluster in majority data. Instance based selection select the relevant data for undersampling. Clustering based Undersampling remove irrelevant data in better manner than existing undersampling technique[8]. To improve the diversity in classifier in stacking, classifier are clustered using clustering algorithm. Selecting classifier from different cluster made the ensemble as expert system. Bagging also done through clustering the data and select the instance from disjoint bag made an bagging an adapative version which is called as A-Bagging[9].Self paced learning used for undersampling algorithm[10].

Ensemble classification can be heterogeneous or homogenous . Heterogeneous ensemble may be formed by selecting better classifier by applying boosting algorithm. This type of ensemble is self configured .It doesn't need expert to select the base classifier because it select automatically[15].Neighbourhood undersampling also used to find the subset of samples from majority class and prediction are done through stacking. Stacking ensemble use Cross validation prediction but Subset and out of subset space also used[14].Instead of static Ensemble, dynamic selection of Ensemble also used[13].Hybrid sampling also possible by combining undersampling and oversampling technique.[11].

Proposed Method

In Imbalanced data classification two type of classes are there one is majority class which is called negative class and another is minority class which is known as positive class. Performances metric are measured for either majority class or overall accuracy .Minority class accuracy are not measured. Proposed architecture consist of two phases

- Tomek link Under Sampling
- Stacking Ensemble Classifier

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2182 – 2190
Received 05 March 2021; Accepted 01 April 2021.

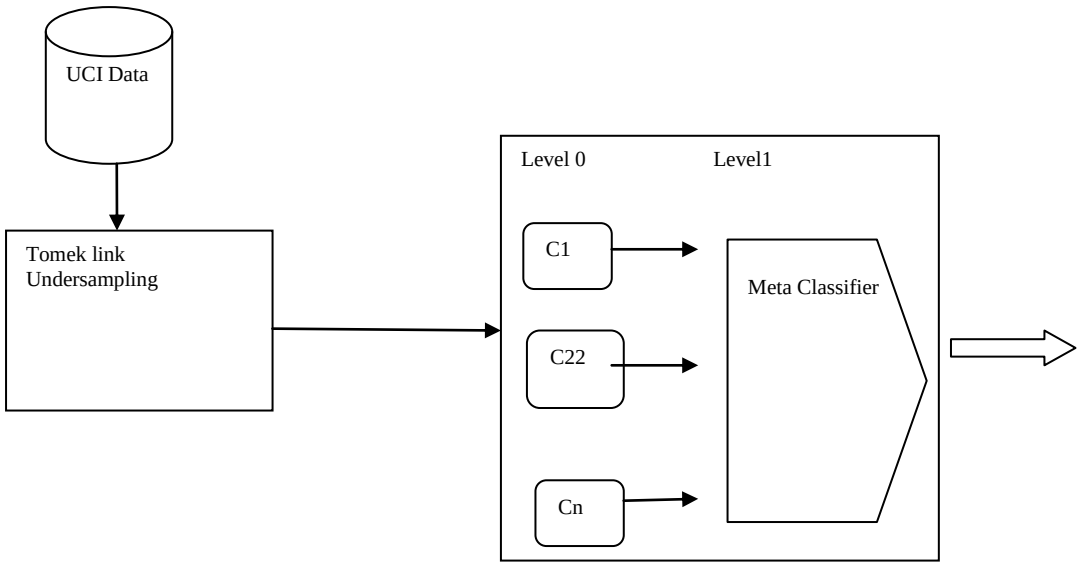

Figure 1.Architecture Diagram

Data set Description

Benchmark Imbalanced dataset are available in UCI Repositry. Both binary class data and multiclass dataset are stored in that repository. Binary class Dataset are taken for Experimental purposes. Among many datasets Medical imbalance data such as Breast-W,Breast cancer hepatisis and Cleveland heart diseases dataset are used for the proposed system. Data set description are given in the table 1.

Table 1:Imbalanced dataset from UCI Repository

	Number	Instances	Attributes	Imbalance Ratio
Hepatisis	2	155	19	1:1.20
Breast -W	2	683	10	1:1.19
Breast Cancer	2	198	32	1:1.32
Heart-Diseases	2	303	13	1:1.50

Tomek Link Undersampling

Tomek link for undersampling is suggested by Ivan Tomek. Using Euclidean distance between the two classes data find closest pair .Tomek link provide linear boundary between pairs. Paired data are closest to each other from two classes. Majority class instances which are closest to minority class instance are either borderline data or noisy data. Usually borderline data and noisy data reduce the accuracy rate for classifier. Tomek link overcome this issue by removing noisy and borderline data

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2182 – 2190
Received 05 March 2021; Accepted 01 April 2021.

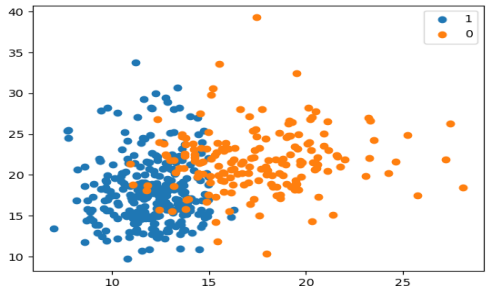

Figure2: Tomek link Undersampling for breast cancer data

Stacked Ensemble classifier

It is Ensemble machine learning model. There are two level of classification has been done Level 0 Individual classifier predicts the result. In level 1 Predicted result are again fed into meta classifier for final prediction For imbalanced data classification state of art classifier give more false positive and false negative. To over come that Ensemble model such as bagging and boosting are suggested.

Ensemble normally take multiple classifier and combine the output using voting technique. Deciding the number of classifier decide the output. Pruning the classifier and find out the best is highly challenging one. Showing diversity in classifier also change the prediction in both the ways.

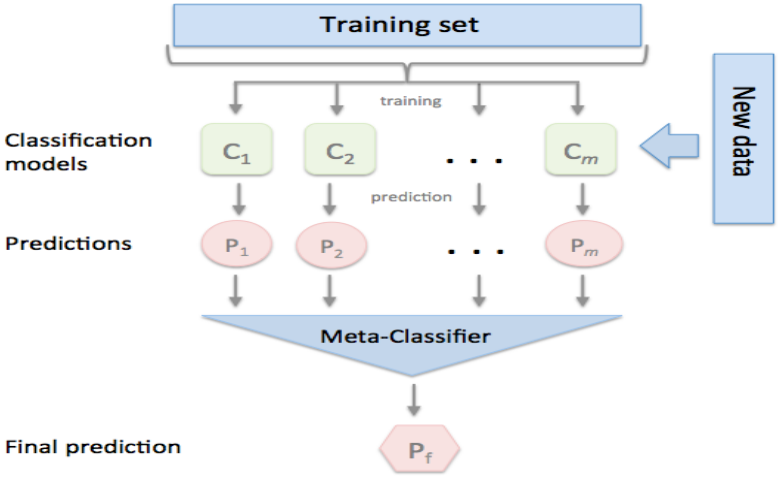

Figure 3 :Stacking Ensemble

Bagging means bootstrapping the dataset and applied to same classifier. Dataset are taken by replacement method. Combining the various output give the final prediction. In case of stacking Ensemble different classifier are used at level 0 that results move to next level prediction. Boosting is lazy learner. Entire dataset is applied to the First level classifier. Next level learner correct the mis predicted label. It will consume time. But the proposed algorithm execute diversity of classifier.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2182 – 2190
Received 05 March 2021; Accepted 01 April 2021.

Level 0 contain Logistic Regression, K-Nearest Neighbour Naïve bayes Decision tree and support vector

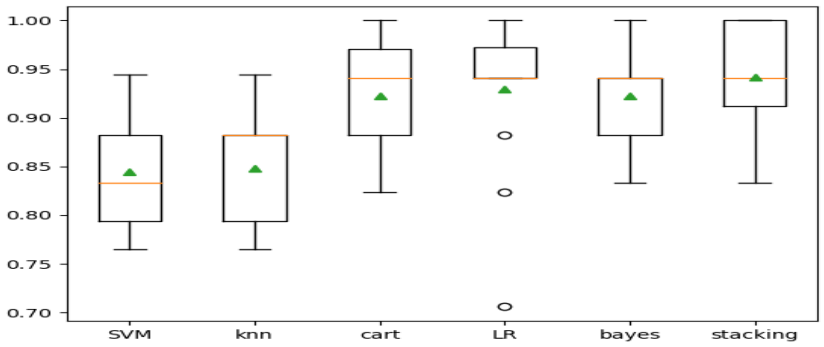

Figure 4 Box plot for comparing stacking Ensemble with individual classifier

machine. Level 1 Logistic regression is used as meta classifier. Applying breast cancer dataset and other data set give accurate prediction. Classifier results are analyzed using box plot. Box or Whisker plot is used to study different experiments on same data. It should not constructed for smaller data set. Fig 4 represent the box plot for breast cancer data set.

Result and Discussion

In binary classification classifier predict the data into any one of the class label either positive or negative. Dataset are divided into training and testing dataset. Algorithm are trained using training dataset and test dataset are predicted .If it predict correctly as positive class then it is True positive and it predicted the negative class correctly it is termed as True negative.

On other side Positive class predicted as negative that is false negative and negative class predicted as positive that is false positive.

Accuracy

Performance measures used to check the correctness of classifier prediction. Most of the classifier appraised using accuracy.

Accuracy can be calculated as

Accuracy=Number of correct prediction/Total number of prediction

Table 2:Accuracy

Dataset	Tomek+ Logistic Regression	Tomek+SVM	Tomek+ Decision Tree	Tomek+ Naïve Bayes	Tomek +KNN	Tomek +Stacking
Breast Cancer	0.93	0.844	0.919	0.848	0.922	0.942
Breast-W	0.93	0.84	0.92	0.92	0.85	0.93
Hepatitis	0.87	0.90	0.90	0.88	0.89	0.91
Heart Diseases	0.85	0.86	0.82	0.84	0.86	0.89

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2182 – 2190
Received 05 March 2021; Accepted 01 April 2021.

Comparing the accuracy of individual classifier with stacked Ensemble classifier reveal that Stacking Ensemble has high accuracy than individual classifier.

Precision

Precision is the ratio of correctly classified positive class sample which is called true positive rate to total number of positive class predicted. Best value is 1 means there is no wrong prediction of positive class

Precision=True positive rate/(true positive rate+false positive rate)

Table 3: Precision Score

Dataset	Tomek+ Logistic Regression	Tomek+ SVM	Tomek+ Decision Tree	Tomek+ Naïve Bayes	Tomek +KNN	Tomek +Stacking
Breast Cancer	0.90	0.80	0.90	0.907	0.907	0.912
Breast-W	0.88	0.84	0.889	0.87	0.91	0.91
Hepatitis	0.85	0.79	0.82	0.81	0.82	0.88
Heart Diseases	0.88	0.80	0.84	0.78	0.86	0.91

Recall

Recall is the ratio of true positive to total positive sample. It is also called sensitivity. Recall measure is important when false negative give great impact.

Recall= (True positive/(True positive+ False negative))

Table 4: Recall Score

Dataset	Tomek+ Logistic Regression	Tomek+ SVM	Tomek+ Decision Tree	Tomek+ Naïve Bayes	Tomek +KNN	Tomek +Stacking
Breast Cancer	0.88	0.80	0.90	0.907	0.907	0.917
Breast-W	0.89	0.81	0.889	0.87	0.91	0.921
Hepatitis	0.85	0.78	0.73	0.70	0.82	0.84
Heart Diseases	0.88	0.80	0.79	0.78	0.86	0.87

From the table 4 Recall score are compared with other classifier. Recall score for various dataset are merely give accurate value.

F1-score

Accuracy is used to measure the correct prediction of class. In imbalanced dataset false positive and false

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2182 – 2190
Received 05 March 2021; Accepted 01 April 2021.

negative rate are more important. To measure accuracy of precision with recall F1 score is used

F1 score=2*((precision *recall)/precision +recall)

From the Table 5 Stacked Ensemble has high value which means reduce the false positive and false negative are
reduced

Table 5: F1 Measure

Dataset	Tomek+ Logistic Regression	Tomek+ SVM	Tomek+ Decision Tree	Tomek+ Naïve Bayes	Tomek +KNN	Tomek +Stacking
Breast Cancer	0.89	0.877	0.892	0.90	0.903	0.91
Breast-W	0.90	0.85	0.902	0.89	0.87	0.92
Hepatisis	0.821	0.83	0.82	0.84	0.88	0.89
Heart Diseases	0.93	0.86	0.91	0.92	0.93	0.95

ROC_AUC score

Receiver operating characteristics (ROC)is used in binary classification. Score can be calculated as true positive rate divided by false positive rate . ROC is used to check the quality of prediction result. it can be illustrated using curve. X-Axis can be taken as false positive rate and Y-axis can be taken as true positive rate. Steepness of the curve decide the quality of the classifier. Area under the steepness can calculated to judge the classifier prediction. Most probable classifier can have roc_auc score as 1.ROCcurve is depicted in Figure 5.

From the table 6 roc_auc score for ensemble sounds good to 0.96 compared to other classifier

Table 6:ROC_AUC score

Dataset	Tomek+ Logistic Regression	Tomek+SVM	Tomek+ Decision Tree	Tomek+ Naïve Bayes	Tomek +KNN	Tomek +Stacking
Breast Cancer	0.912	0.93	0.91	0.93	0.87	0.95
Breast-W	0.90	0.92	0.94	0.88	0.90	0.96
Hepatitis	0.86	0.87	0.90	0.92	0.87	0.91
Heart Diseases	0.95	0.90	0.867	0.92	0.93	0.94

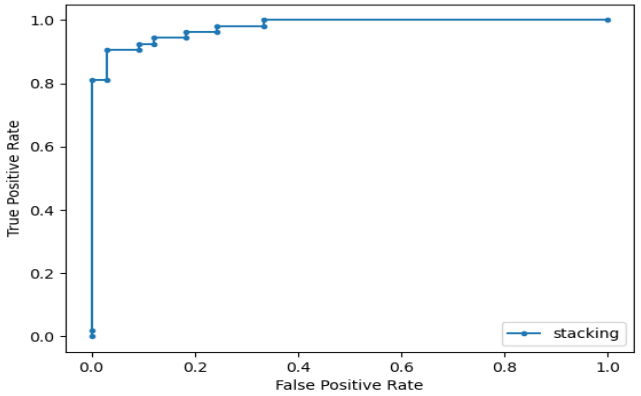

Figure5 ROC curve for stacking Ensemble of Breast cancer dataset

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2182 – 2190
Received 05 March 2021; Accepted 01 April 2021.

Conclusion

Imbalance classification problem is one of current research area in the field of Data Mining. Predicting the class label for the data is referred as classification. Classification results as skewed distribution for some dataset. Prediction Accuracy for imbalanced dataset has wide variety of application in medical ,weather prediction and financial area. To improve the accuracy of imbalanced dataset by reducing the false positive rate and false negative rate. Imbalanced dataset such as breast cancer, heart diseases ,hepatitis are taken from UCI Repository. Undersampling technique are used to balance the data by reducing the samples in majority class. Tomek link undersampling techniques are used to find the distance between two points between majority and minority class. Closest points between two classes are removed which may be noisy or borderline data. After sampling balanced data are load to the stacking Ensemble. It consists of two levels. Level 0 contain many individual classifier such as naïve bayes, SVM, etc., Output of level 0 is feed to level classifier for final prediction. Datasets are applied to base classifier such as Logistic Regression, K-Nearest Neighbour and Naïve bayes. Performance measures such as accuracy, precision recall f1 score and roc_auc score are compared for both stacking Ensemble and base classifier. For imbalanced dataset ROC_AUC score is the important metric Stacking Ensemble give

References

[1] . Salunkhe, U. R., & Mali, S. N. (2016). Classifier ensemble design for imbalanced data classification: a hybrid approach. *Procedia Computer Science*, *85*, 725-732.

[2] Xia, Y., Liu, C., Da, B., & Xie, F. (2018). A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Systems with Applications*, *93*, 182-199.

[3] . D. Liang, C. Tsai, A.-J. Dai, W. Eberle, A novel classifier ensemble approach for financial distress prediction, Knowl. Inf. Syst. 54 (2018) 437–462

[4] .Maji, S., & Arora, S. (2019). Decision tree algorithms for prediction of heart disease. In *Information and Communication Technology for Competitive Strategies* (pp. 447-454). Springer, Singapore.

[5] .Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, *16*, 100203.

[6] Shen, F., Zhao, X., Li, Z., Li, K., & Meng, Z. (2019). A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation. *Physica A: Statistical Mechanics and its Applications*, *526*, 121073.

[7] . Sundar, R., & Punniyamoorthy, M. (2019). Performance enhanced Boosted SVM for Imbalanced datasets. *Applied Soft Computing*, *83*, 105601.

[8] Tsai, C. F., Lin, W. C., Hu, Y. H., & Yao, G. T. (2019). Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. *Information Sciences*, *477*, 47-54

[9] Agarwal, S., & Chowdary, C. R. (2020). A-Stacking and A-Bagging: Adaptive versions of ensemble learning algorithms for spoof fingerprint detection. *Expert Systems with Applications*, *146*, 113160.

[10] Wang, Q., Zhou, Y., Zhang, W., Tang, Z., & Chen, X. (2020). Adaptive sampling using self-paced learning for imbalanced cancer data pre-diagnosis. *Expert Systems with Applications*, *152*, 113334.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2182 – 2190
Received 05 March 2021; Accepted 01 April 2021.

[11]. Gao, X., Ren, B., Zhang, H., Sun, B., Li, J., Xu, J., ... & Li, K. (2020). An ensemble imbalanced classification method based on model dynamic selection driven by data partition hybrid sampling. *Expert Systems with Applications*, *160*, 113660.

[12]. Du, X., Li, W., Ruan, S., & Li, L. (2020). CUS-heterogeneous ensemble-based financial distress prediction for imbalanced dataset with ensemble feature selection. *Applied Soft Computing*, *97*, 106758.

[13].Hou, W. H., Wang, X. K., Zhang, H. Y., Wang, J. Q., & Li, L. (2020). A novel dynamic ensemble selection classifier for an imbalanced data set: An application for credit risk assessment. *Knowledge-Based Systems*, *208*, 106462.

[14]. Seng, Z., Kareem, S. A., & Varathan, K. D. (2020). A Neighborhood Undersampling Stacked Ensemble (NUS-SE) in imbalanced classification. *Expert Systems with Applications*, 114246.

[15].Kadkhodaei, H. R., Moghadam, A. M. E., & Dehghan, M. (2020). HBoost: A heterogeneous ensemble classifier based on the Boosting method and entropy measurement. *Expert Systems with Applications*, 113482.