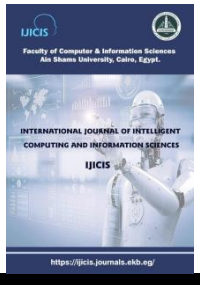
IJICIS, Vol.23, No.3, 83-94
DOI: 10.21608/ijicis.2023.216833.1278

International Journal of Intelligent Computing and Information Sciences

https://ijicis.journals.ekb.eg/



SMOTE-RUS: COMBINED OVERSAMPLING AND UNDERSAMPLING TECHNIQUE TO CLASSIFY THE IMBALANCED AUTISM SPECTRUM DISORDER DATASET

Eman Ismail*

Information Systems Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
emanismail@cis.asu.edu.eg

Walaa Gad

Information Systems Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
walaagad@cis.asu.edu.eg

Mohamed Hashem

Information Systems Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
mhashem@cis.asu.edu.eg

Received 2023-06-10; Revised 2023-06-10; Accepted 2023-07-29

Abstract: *The imbalanced distribution of classes is a common issue in almost classification problems. Therefore, we must be familiar with class-imbalanced techniques to handle this problem. Autism spectrum disorder(ASD) disease affects the development of the brain. Therefore, patients with autism have some limitations to interact with others on the social level. So, it is necessary to predict the genes related to ASD for early diagnosis and treatment. Recent studies utilize different machine learning techniques to predict ASD genes that suffer from the imbalanced ASD dataset problem. In this paper, recent ASD gene prediction models are utilized to compare different techniques influence using undersampling and oversampling algorithms on the model performance. Moreover, a new combined technique(SMOTE-RUS) is proposed using Synthetic Oversampling Technique(SMOTE) and random undersampling(RUS) technique to solve the imbalanced dataset problem. SMOTE-RUS is used to build an effective model to predict ASD genes. The proposed technique results prove that it is effective to get a more robust gene prediction model. Moreover, it outperforms other models using a single resampling technique.*

Keywords: *Oversampling, Undersampling, SMOTE, Gene prediction, Class imbalance problem.*

1. Introduction

The imbalanced dataset remains one of the most machine learning (ML) challenges. Although ML has a great effect in most factual applications, it is necessary to be aware of the useful techniques to learn from the imbalanced dataset. Imbalanced classification means that their classes have a skewed class distribution, as there is a big difference in the size of class samples. When samples size of one class elevates samples size of other class, it means there is an imbalance in the class distribution. The

***Corresponding Author**: Eman Ismail

Information Systems Department, Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt

Email address: emanismail@cis.asu.edu.eg

traditional classification techniques failed to detect the correct class, as they are built to use in a balanced dataset. Although, most real-time applications target the class that is considered the minority one. Most traditional techniques will bias toward the majority category and discard the minority one if the dataset is unbalanced. Recently diverse techniques were proposed by researchers to handle the imbalanced dataset problem.

There are different approaches used to learn from the imbalanced dataset which are data-based approaches, algorithm-based approaches, and hybrid form approaches [1]. The data-based approaches are worked in the preprocessing steps before the classification techniques. It is categorized into three different classes, undersampling techniques [2], oversampling techniques [3], and a hybrid technique of under and oversampling. The algorithm-based approaches depend on the classification algorithm itself as they decrease the sensitivity of the classification algorithm towards the majority class, which is called "Cost-sensitive techniques". The hybrid form approaches are integration between approaches of data and algorithm techniques together.

In this work, autism spectrum disorder (ASD) gene prediction is our target to early detect people with autism and find innovative treatment methods. ASD is a disease that is related to the neurodevelopment of the brain, which affects the social behavior of the patient. Autism is diagnosed in a large proportion through the apparent symptoms of the disease due to the lack of genes associated with autism that have been discovered so far. Therefore, the discovery of many genes associated with the disease is very important for proper diagnosis and better treatment. Most studies of gene prediction use machine learning (ML) techniques which are efficient enough to predict disease genes. However, there is still a common problem using ML techniques in the case of imbalanced datasets. Most bioinformatics and medical domain suffer from the imbalanced dataset problem as in the case of autism dataset. Therefore, we propose a new hybrid technique to handle the imbalanced ASD dataset problem and form the most powerful model. The proposed model predicts the largest number of ASD genes. This technique combines oversampling and undersampling techniques to take advantage of both techniques making a balanced dataset.

The rest of the paper sections are arranged as follows: the state of art methods that are used to solve the imbalanced dataset problem and the recent ASD gene prediction models are described in section 2. Section 3 describes the proposed resample technique using oversampling and undersampling techniques and describes the proposed gene prediction model. Section 4 concludes all experimental results, and discussion. Finally, the conclusion of this article is presented in section 5.

2. Related Works

Diverse methods are used to solve the imbalanced dataset problem [1], which considers data-based approaches. Data-based approaches are applied in the preprocessing of the data such as oversampling methods and undersampling methods [2]. Oversampling methods increase minority class samples either by repeating some data samples from the minority class or creating new instances using different techniques. The advantage of oversampling techniques is that there is no loss of information as the original samples of the dataset keep as it is besides the new samples added that may be useful. The oversampling techniques disadvantage is that they sometimes take a long time to execute compared to undersampling techniques. Moreover, in some cases, oversampling techniques caused an overfitting problem if they duplicated some samples from the dataset. There are different oversampling techniques

such as random oversampling [3], smote [4], Borderline smote [5], k-means smote [6], Adaptive Synthetic (ADASYN) Sampling Approach [7], and SVM smote [8]. Random oversampling is the simplest technique as it randomly duplicates some samples from the majority class. ¸SMOTE is the most effective oversampling technique, as it creates new synthetic samples from the minority samples. Borderline smote, k-means smote, Adaptive Synthetic (ADASYN) Sampling Approach, and SVM smote are different variations of SMOTE algorithm that are all effective and have similar performance. Undersampling techniques are simple and fast, but they may lose some information as they remove some samples from the majority class to make a balanced dataset. ML methods are used to construct a gene predictive model as binary classification problem that has positive and negative samples. Support Vector Machine (SVM) [9], K-Nearest Neighbor (KNN) [10], Naïve Bayes (NB) [11], Artificial Neural Network (ANN) [12], and Decision Trees (DT) [13] are the most used machine learning methods in gene prediction models. Moreover, ensemble learning techniques [14, 15, 16] are used in gene prediction models which combine single classifiers to form a more robust model. In [17], they proposed a ML-based model to predict ASD risk genes. They trained their model using gene expression data of brain development. They used the bayes network to predict the risk ASD genes and applied discretization on the data as a preprocessing step to enhance the model performance. Moreover, authors in [18] use different machine learning classifiers to test if the child was susceptible to affect with autism in his early stages using SVM, KNN, Naïve Bayes, and Logistic regression. Logistic regression got the highest accuracy using their selected ASD dataset.

3. Proposed Model

The proposed prediction model framework is explained in Figure.1. It includes several processes to build an effective predictive model. It uses SFARI gene database https:// gene. sfari org/ to evaluate the performance of the presented model. SFARI holds all candidates' genes related to autism and each gene has a score representing his correlation with ASD disease. It has several categories, only categories one, two, three, and four are included in evaluating the proposed model. Moreover, the genes that are in the syndrome category and included in any of one, two, three, and four are included in the evaluation. Categories number one and number two have the highest score as they are most relevant to ASD, and categories three and four have lower scores than categories one and two. Each gene is annotated using Gene Ontology (GO) [19, 20]. GO is represented as a hierarchical structure, as the nodes represent the gene term, and the associations between terms are represented as graph edges. It is divided into three paths; the first path represents terms participants in any biological processes which are responsible for the living of the cell beginning from its configuration till its final product. The second path represents the molecular function which includes all gene terms that represent the activities of the gene product regardless of where or how these actions are done. The third path represents the cellular component which contains the gene terms that participate in the building of the structure of the cell. In this paper, the gene terms of all biological processes are included in the analysis of the proposed model.

Sequentially, we construct the gene functional similarity matrix that contains all the candidates' genes and semantic similarity values between them. Therefore, a new measure of similarity method [21] is used to calculate the similarity among the genes. This function is called HGS that is a hybrid method between Wang's [22] method and information content (IC) methods. It is depending on Wang's method and acts as IC methods. It takes the benefits from IG methods using the number of the term children rather than searching for the term in a large corpus file.

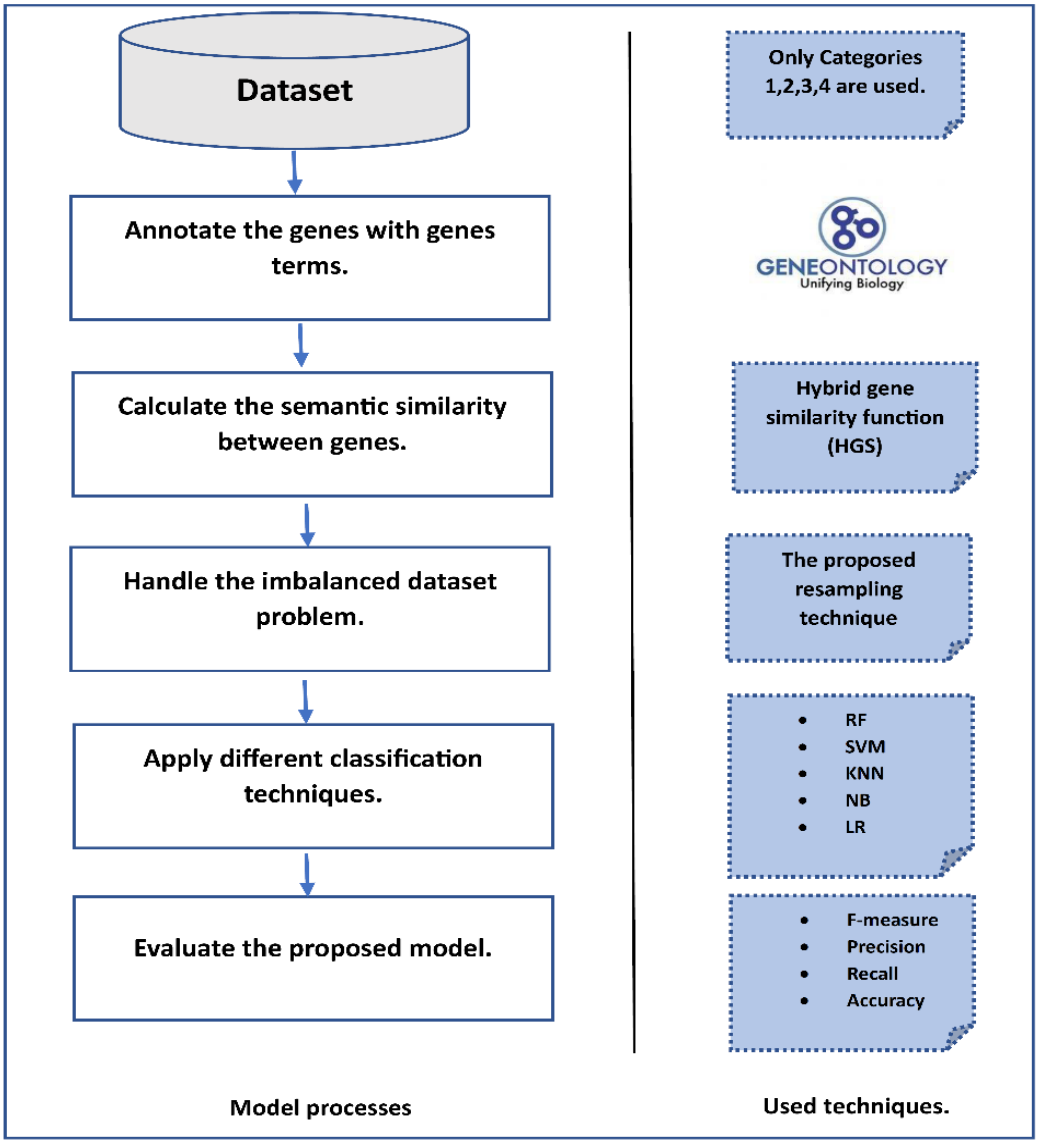


Figure 1: Gene prediction Framework

3.1 SMOTE-RUS Resampling Technique

SFARI dataset faces the imbalanced dataset problem. Therefore, to avoid the negative impact of this problem on the classification process. We propose a new combined resampling technique "SMOTE-RUS" as a new data-based approach. SMOTE-RUS is Synthetic Minority Oversampling TEchnique- Random Undersampling. SMOTE [4] is combined with RUS to avoid the disadvantages of the RUS technique. The following paragraphs explain RUS and SMOTE in detail.

Random Undersampling (RUS): is a very straightforward technique that randomly eliminates some data samples from majority class to make a balanced dataset. This technique is applied before the classification process, it is easy and fast but has some limitations. RUS solves the imbalanced problem, but it can remove some important samples from the majority class which may have useful

information reflected on the performance of the classification algorithms. Therefore, RUS is not a suitable technique in some cases. Rus technique steps are as follows:

- Load SFARI dataset and define the size of majority and minority samples.
- Define the resampling percentage to discard some data samples from majority samples.
- Choose a random sample to remove it.
- Iterate the prior steps until we reach the identified percentage.

Synthetic Minority Oversampling TEchnique (SMOTE): creates a new synthetic sample from the minority category. It is different from other oversampling techniques as they duplicate the minority samples to make a balanced dataset. SMOTE used the KNN algorithm to identify the Nearest Neighbor (NN) to the randomly selected sample and generate the synthetic samples. The general algorithm steps of SMOTE are identified as follows:

- Load SFARI genes data and define the count of majority and minority data samples.
- Define the resampling percentage to generate some samples in the minority category.
- Choose a random sample from the minority category and use KNN [23] algorithm to find its nearest neighbor.
- Choose one of its nearest neighbors (NN) then find the variation between the selected random sample and this NN.
- Multiply this difference with a random number between (0,1).
- Combine the result value after multiplication with the chosen random sample to form the new synthetic sample.
- Repeat the previous four steps until we reach the given resampling percentage.

SMOTE-RUS Technique: is proposed as a new combined technique to handle the imbalanced dataset problem. It combines SMOTE with a random undersampling technique to avoid its disadvantages. The detailed steps of the SMOTE-RUS are shown in Figure. 2. In the first step, the count of minority and majority data samples is counted to determine the required balanced ratio to handle the imbalanced dataset. After that, the minority samples are oversampled using SMOTE. Synthetic samples are created from the minority class samples until we reach 100% percentage for a balanced ratio. The KNN algorithm is used to calculate the five Nearest Neighbor (NN) of the selected minority sample to create new synthetic samples added to the minority category. The new synthetic samples are created until we reach the required balanced ratio. Sequentially, the majority class is undersampled using RUS. RUS deletes random data samples from the majority category till we reach the required balance ratio using Eq. (1).

$$\text{precUnder} = (\text{Num Of Positive Instances}/\text{Num Of Negative Instances} * 100) \quad (1)$$

After, handling the imbalanced dataset problem and constructing the matrix of functional gene similarity. Different classification techniques are used to predict ASD genes utilizing different classifiers such as RF [24], SVM [25], KNN [23], and NB [26]. Moreover, different assessment measures are used to assess the proposed model using these classifiers such as REcall, PRecision, F-Measure, and ACCuracy using Cross-Fold validation method [27].

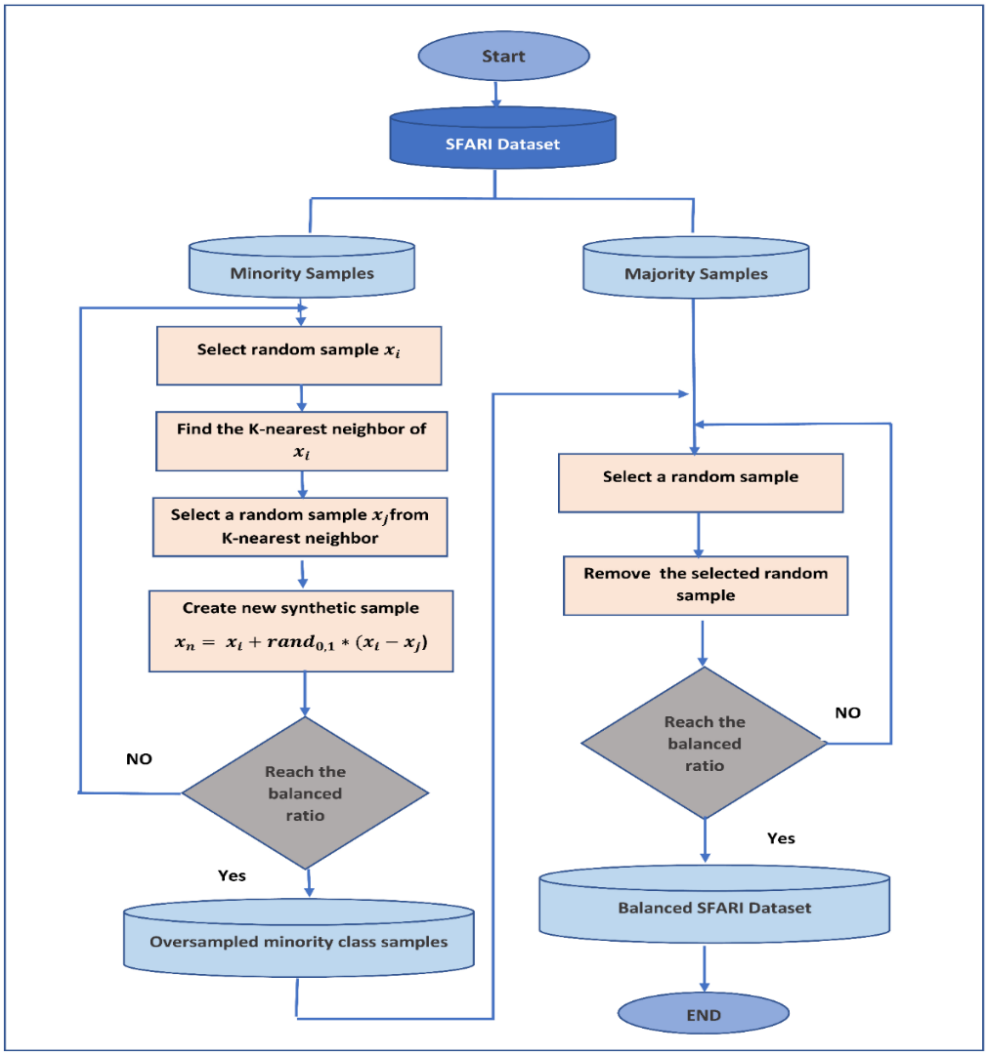


Figure 2:The flowchart of SMOTE-RUS

4. Experimental Results

4.1 Databases Description

SFARI gene dataset is used to assess the presented prediction model using the proposed combined resampling technique. It contains all candidates' genes that are related to ASD. It is classified into two categories, the first one is called Highest Confidence Genes (HCG), and the second one is called Lowest Confidence Genes (LCG). SFARI genes are counted as 990 genes, 82 genes from them are considered HCG, and the LCG are counted as 506 genes. The rest of the genes that have no score relate them to ASD are excluded from the analysis of the proposed model. Moreover, more genes are involved in the analysis result from [28] which are 1189 non-mental genes that added to class "non-ASD" as negative samples.

4.2 Evaluation Metric

Stratified cross-fold validation is utilized to assess the proposed model, which splits the dataset into equal folds. The assessment is done till fold five. In each iteration, four folds are employed in training and the rest one for testing. Also, in each iteration, a different fold is used for testing. The performance of the model is assessed using various performance metrics, which are Precision in Eq.2, Recall in Eq.3, F-Measure in Eq.4, and Accuracy in Eq.5.

$$PRecision = Tp/(Tp + FP) \quad (2)$$

$$REcall = Tp/(Tp + FN) \quad (3)$$

$$F - Measure = (2 * REcall * PRecision)/(REcall + PRecision) \quad (4)$$

$$ACCuracy = (Tp + TN)/(Tp + TN + FP + FN) \quad (5)$$

True positive (Tp) is a value reflected that the model predicts the right positive "ASD" category. True Negative (TN) is a value reflected that the model predicts the right negative "non-ASD" category. False Positive (FP) is a value reflected that the model predicts the fault positive "ASD" class. False Negative (FN) is a value reflected that the model predicts the fault negative "non-ASD" class.

4.3 Results

In [18], they build two different forms for the matrix of functional gene similarity. In the first form, they use HCG and the 1189 non-mental genes [28], and in the second form, they use HCG, LCG, and the non-mental genes to assess the performance of their model. They build a predictive model to predict ASD genes using various function of semantic similarity measures such as Relevance [29], Wang's [22], Resnik [30], and propose HGS function [23] which gained the highest performance. Their model used the trivial resampling technique to handle the imbalanced SFARI dataset problem. In this work, we build our predictive model using our proposed combined technique "SMOTE-RUS" to fix the imbalanced dataset problem. The proposed model performance is compared against the HEC model [21] using the first form of gene functional similarity matrix as its results have the highest performance [20]. Figure.3 shows the results of the presented two techniques to fix the imbalanced SFARI dataset in our predictive model using different classifiers in terms of precision.

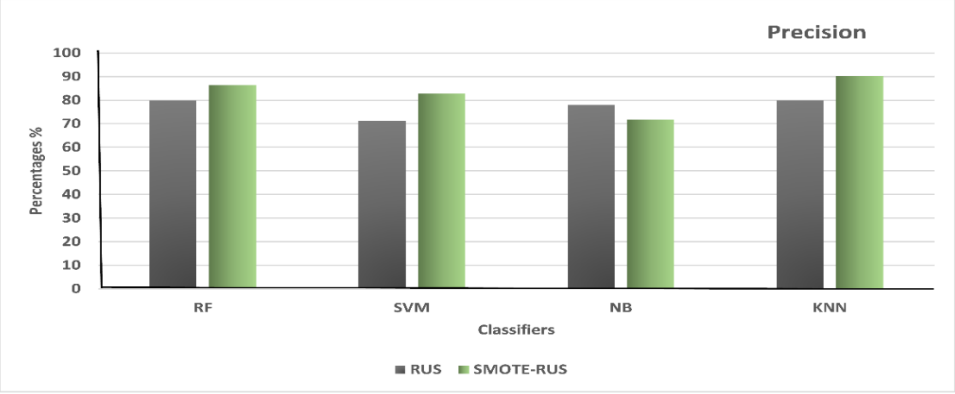


Figure 3: Different classifiers performance using different resampling techniques in terms of percision

These techniques are random undersampling (RUS) and the proposed SMOTE-RUS technique and different classifiers are employed to assess the proposed prediction model such as SVM, RF, NB, and KNN. Moreover, Figure.4, 5, and 6 show the results of the presented model using the proposed SMOTE-RUS against RUS using REcall, F-Measure, and ACCuracy. The proposed SMOTE-RUS outperforms the RUS technique, which reaches the proposed model an accuracy around 88% using KNN rather than 79% using RUS. This improvement indicates that the proposed SMOTE-RUS is valuable in handling the imbalanced problem reflected in enhancing the prediction of ASD genes.

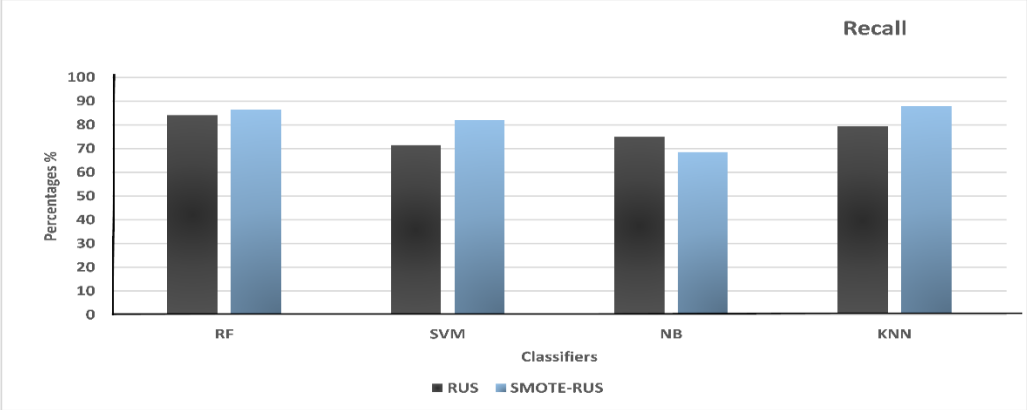


Figure 4: Different classifiers performance using various techniques of resampling with recall

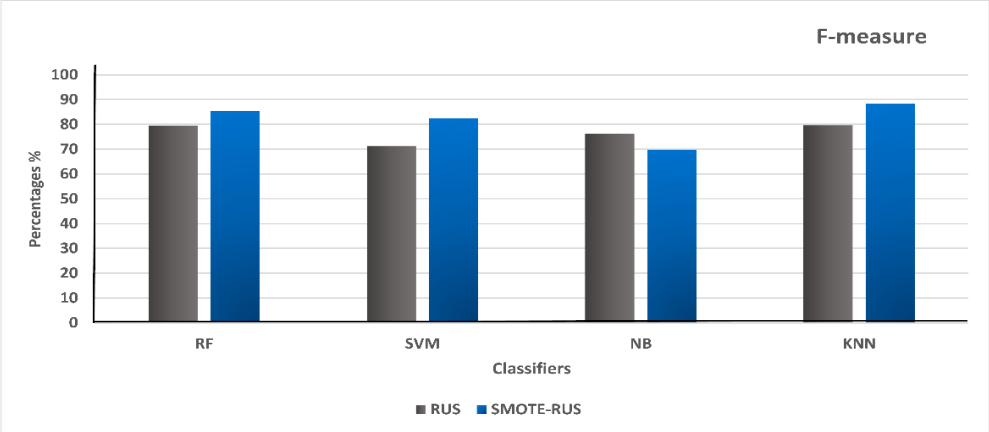


Figure 5: Different classifiers performance using different techniques of resampling with f-measure

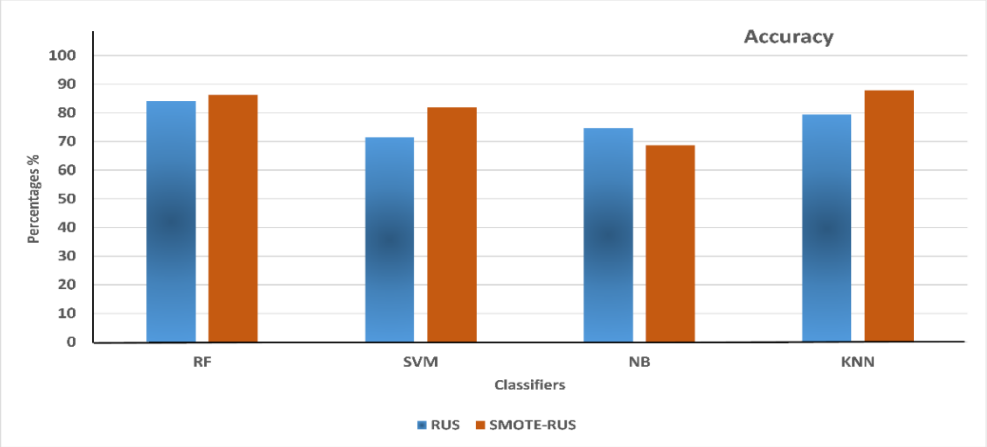


Figure 6: Different classifiers performance using different techniques of resampling with accuracy

Moreover, ensemble ML techniques are used to assess the presented model using SMOTE-RUS to compare the efficiency of our predictive model against the HEC-ASD predictive model [21]. The first technique is Adaboost [31], which is the trivial method for boosting techniques that attempt to build a strong model. Fig.7 shows the proposed model results using Adaboost-RUS and Adaboost-SMOTE-RUS. The Adaboost-SMOTE-RUS increases the performance around 0.5 rather than using RUS. For more enhancement, gradient boosting (GB) [32] is used, which is another form of boosting technique that is most updated and more effective than Adaboost. Fig.8 shows a comparison between our proposed model using gradient boosting technique and HEC-ASD [21] based on GB technique. The proposed model is based on GB using SMOTE-RUS get the highest performance around 90.5% compared to HEC-ASD [21] which reached an accuracy 87.5% using gradient boosting and RUS techniques.

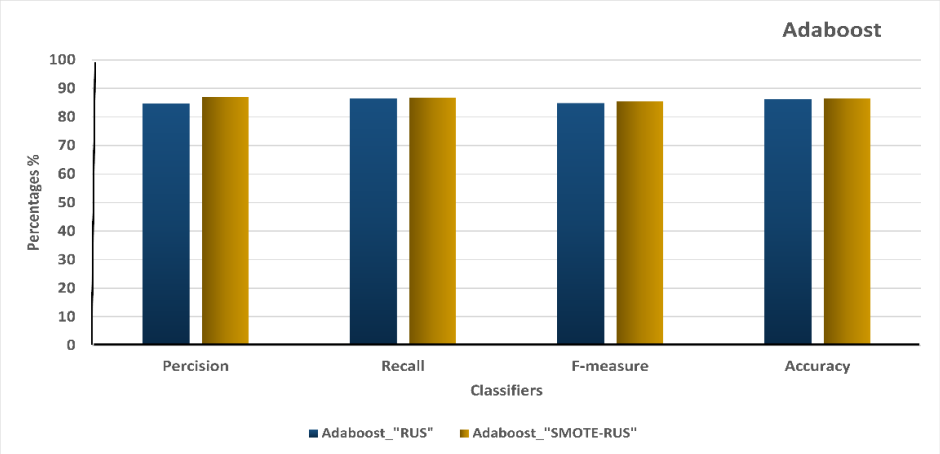


Figure 7: Results of Adaboost classification technique using different resampling techniques

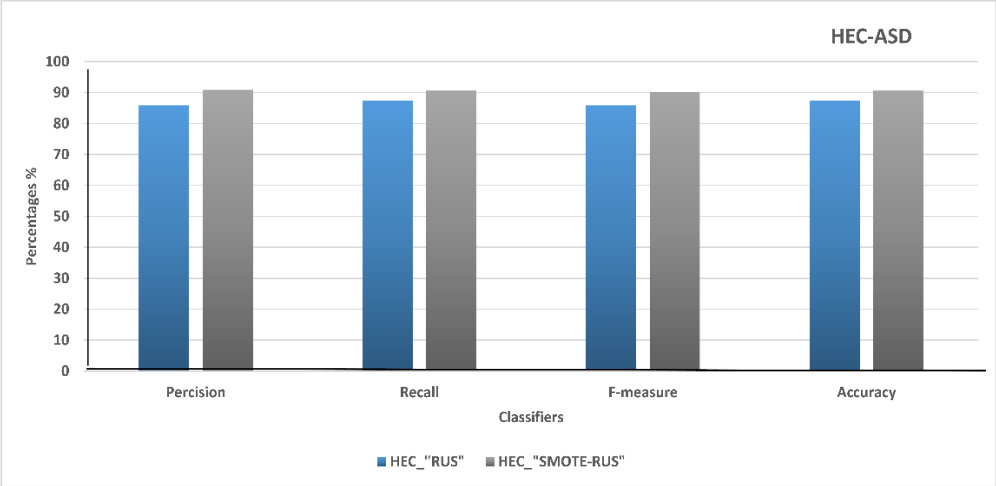


Figure 8: Comparison between the proposed model and HEC-ASD model

4. Conclusion

Autism Spectrum Disorder (ASD) disease is a complex disease which is considered the most prevalent disease among children. Early diagnosis and treatment will help in treating these children and not exacerbating the symptoms of their disease. Therefore, we must be aware of the disease's causes. In this article, we build a predictive model to identify the genes that cause ASD. There are few genes that caused ASD are predicted. Therefore, we propose a new combined technique SMOTE-RUS to handle the imbalanced dataset problem of ASD as the majority class is the "non-ASD" genes that affect the predictive model performance. The proposed model uses GO to annotate the genes and uses an effect method HGS to calculate the semantic similarity between the genes. Moreover, diverse classifiers are used to assess the efficiency of the model as SVM, RF, NB, and KNN. Ensemble ML techniques are used as Adaboost and Gradient Boosting GB technique to build a more robust predictive model. The presented model results using GB and SMOTE-RUS to handle the imbalanced dataset problem get the highest performance accuracy around 90.5% compares to other techniques that reached an accuracy of 87.5%. Therefore, the proposed predictive model is effective in predicting the genes that caused ASD. But it has some limitations as there are a few numbers of genes that do not have any annotation in GO so we ignore them in the analysis process. Therefore, in the future work, more improvements can be made by integrating some other annotation resources with GO to get a more powerful prediction model.

References

1. Batista, G.E., Prati, R.C. and Monard, M.C., 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), pp.20-29.
2. Mohammed, R., Rawashdeh, J. and Abdullah, M., 2020, April. Machine learning with oversampling and undersampling techniques: overview study and experimental results. In *2020 11th international conference on information and communication systems (ICICS)* (pp. 243-248). IEEE..

3. Moreo, A., Esuli, A. and Sebastiani, F., 2016, July. Distributional random oversampling for imbalanced text classification. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 805-808).
4. Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, pp.321-357.
5. Han, H., Wang, W.Y. and Mao, B.H., 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *Advances in Intelligent Computing: International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I 1* (pp. 878-887). Springer Berlin Heidelberg.
6. Douzas, G., Bacao, F. and Last, F., 2018. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Information Sciences*, *465*, pp.1-20.
7. He, H., Bai, Y., Garcia, E.A. and Li, S., 2008, June. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322-1328). IEEE.
8. Wang, H.Y., 2008, June. Combination approach of SMOTE and biased-SVM for imbalanced datasets. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* (pp. 228-231). IEEE.
9. Keerthikumar, S.H.I.V.A.K.U.M.A.R., Bhadra, S.A.H.E.L.Y., Kandasamy, K.U.M.A.R.A.N., Raju, R.A.J.E.S.H., Ramachandra, Y.L., Bhattacharyya, C.H.I.R.A.N.J.I.B., Imai, K.O.H.S.U.K.E., Ohara, O.S.A.M.U., Mohan, S.U.J.A.T.H.A. and Pandey, A.K.H.I.L.E.S.H., 2009. Prediction of candidate primary immunodeficiency disease genes using a support vector machine learning approach. *DNA research*, *16*(6), pp.345-351.
10. Xu, J. and Li, Y., 2006. Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics*, *22*(22), pp.2800-2805.
11. Lage, K., Karlberg, E.O., Størling, Z.M., Olason, P.I., Pedersen, A.G., Rigina, O., Hinsby, A.M., Tümer, Z., Pociot, F., Tommerup, N. and Moreau, Y., 2007. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature biotechnology*, *25*(3), pp.309-316.
12. Sun, J., Patra, J.C. and Li, Y., 2009, June. Functional link artificial neural network-based disease gene prediction. In *2009 International Joint Conference on Neural Networks* (pp. 3003-3010). IEEE.
13. Adie, E.A., Adams, R.R., Evans, K.L., Porteous, D.J. and Pickard, B.S., 2005. Speeding disease gene discovery by sequence based candidate prioritization. *BMC bioinformatics*, *6*, pp.1-13.
14. Yang, P., Li, X., Chua, H.N., Kwoh, C.K. and Ng, S.K., 2014. Ensemble positive unlabeled learning for disease gene identification. *PloS one*, *9*(5), p.e97079.
15. Alkuhlani, A., Gad, W., Roushdy, M. and Salem, A.B.M., 2022. Pustackngly: positive-unlabeled and stacking learning for n-linked glycosylation site prediction. *IEEE Access*, *10*, pp.12702-12713.
16. Alkuhlani, A., Gad, W., Roushdy, M. and Salem, A.B.M., 2021. Intelligent techniques analysis for glycosylation site prediction. *Current Bioinformatics*, *16*(6), pp.774-788.
17. Gök, M., 2019. A novel machine learning model to predict autism spectrum disorders risk gene. *Neural Computing and Applications*, *31*(10), pp.6711-6717.
18. Vakadkar, K., Purkayastha, D. and Krishnan, D., 2021. Detection of autism spectrum disorder in children using machine learning techniques. *SN Computer Science*, *2*, pp.1-9.
19. Yu, G., 2020. Gene ontology semantic similarity analysis using GOSemSim. *Stem Cell Transcriptional Networks: Methods and Protocols*, pp.207-215.

20. Ismail, E., Gad, W. and Hashem, M., 2021, December. Predicting of autism spectrum disorder using gene ontology. In *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS)* (pp. 442-447). IEEE.
21. Ismail, E., Gad, W. and Hashem, M., 2022. HEC-ASD: a hybrid ensemble-based classification model for predicting autism spectrum disorder disease genes. *BMC bioinformatics*, *23*(1), p.554.
22. Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S. and Chen, C.F., 2007. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, *23*(10), pp.1274-1281.
23. Guo, G., Wang, H., Bell, D., Bi, Y. and Greer, K., 2003. KNN model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings* (pp. 986-996). Springer Berlin Heidelberg.
24. Rigatti, S.J., 2017. Random forest. *Journal of Insurance Medicine*, *47*(1), pp.31-39.
25. Suthaharan, S. and Suthaharan, S., 2016. Support vector machine. *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, pp.207-235.
26. Webb, G.I., Keogh, E. and Miikkulainen, R., 2010. Naïve Bayes. *Encyclopedia of machine learning*, *15*, pp.713-714.
27. Fushiki, T., 2011. Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, *21*, pp.137-146.
28. Krishnan, A., Zhang, R., Yao, V., Theesfeld, C.L., Wong, A.K., Tadych, A., Volfovsky, N., Packer, A., Lash, A. and Troyanskaya, O.G., 2016. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nature neuroscience*, *19*(11), pp.1454-1462.
29. Pesquita, C., Faria, D., Falcao, A.O., Lord, P. and Couto, F.M., 2009. Semantic similarity in biomedical ontologies. *PLoS computational biology*, *5*(7), p.e1000443.
30. Resnik, P., 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of artificial intelligence research*, *11*, pp.95-130.
31. Schapire, R.E., 2013. Explaining adaboost. *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pp.37-52.
32. Natekin, A. and Knoll, A., 2013. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*, p.21.