International Journal of Computer Sciences and Engineering
Vol.**11**, Special Issue.**1**, pp.**62-65**, November **2023**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Paper

Handling Imbalanced Heart Disease Data and Explaining the Factors

Sandip Das[1*], Gairik Sajjan[2], Arkajyoti Poddar[3], Tamojit Dasgupta[4], Sayani Patty[5], Debmitra Ghosh[6]

[1,2,3,4,5,6]Dept. of Computer Science and Engineering, JIS University, Kolkata, India

**Corresponding Author: dsandip2233@gmail.com*

Abstract: Heart disease is one of the most serious and life threatening problems. If predicted beforehand, many lives can be saved. But, the problem is that medical datasets are highly imbalanced, which leads machine learning algorithms to perform poorly on the minority class. Which in terms leads to wrong predictions. In healthcare it is highly risky to predict something wrongly, because, people's lives are on stake. The ratio of minority and majority class data should be 1:1, or near about equal, in order to get a good result. Synthetic Minority Oversampling TEchnique(SMOTE) is one such oversampling technique that makes it come true, which is used in this work. In addition we have used eXplainable AI(XAI) to better visualise the predictions. We have used LIME (Local Interpretable Model-agnostic Explanation) and SHAP (Shapely Additive Explanations) algorithms to understand the contributions of features towards the predictions.

Keywords: Heart Disease, SMOTE, Machine Learning, Explainable AI, LIME, SHAP

1. Introduction

An imbalanced dataset is one in which the number of samples in each class is not equal. For instance, a dataset that contains data on the occurrence of a rare disease might have only a few positive examples and many negative ones, resulting in class imbalance.This problem can have several negative consequences, including biased model predictions, low sensitivity or recall, and a lack of generalizability. When training on imbalanced data, the model might learn to predict the majority class, resulting in poor performance on the minority class.

In the medical field, heart diseases are one of the major concerns. Medical teams are continuously working on this field and trying to find a predictive solution. Heart disease data is highly imbalanced in the medical field, making it very difficult to get good predicted results. We observe that few machine learning models may give high accuracy but in terms of precision and confusion matrix, this is highly unacceptable. As it is data from the medical field, it is highly risky to predict wrongly; it can cause both mental and social illness. For this reason, we cannot train any machine learning model with imbalanced class data. We have to balance the classes. There are several approaches to address imbalanced medical datasets, including resampling, cost-sensitive learning, and ensemble methods. Resampling involves either oversampling the minority class, undersampling the majority class, or a combination of both. Oversampling can be done by duplicating examples or generating synthetic examples, while undersampling involves removing examples from the majority class. Cost-sensitive learning involves assigning different costs to different types of errors. For instance, a false negative might be more costly than a false positive in medical diagnosis, and the model can be trained to minimize the overall cost. Ensemble methods involve combining several models to improve performance. For instance, one can train several models on different subsets of the data, and then combine their predictions.

In our working dataset, we used an oversampling method called SMOTE (Synthetic Minority Oversampling Technique). After balancing the dataset, we got good accuracy as well as good values of recall, precision, and confusion matrix.

Machine learning models are like black boxes; we give an input and get an output. Explainable AI is a type of Artificial Intelligence that helps people understand how a machine learning model makes decisions. It can help us understand why a model did something, so that we can trust it and use it better. Basically, it shows the insights of a machine learning model. Using XAI, we can analyze which attributes contribute how much and which attribute is more responsible for predicting such as.

2. Related Work

Handling Imbalanced Healthcare Data with Supervised and Unsupervised Methods: A Systematic Literature Review by Deldar, Mahdavi, and Mohammadzadeh. In this study[1], the authors review the existing literature on handling imbalanced healthcare data using both supervised and unsupervised methods.

Handling Imbalanced Data in Healthcare: A Systematic Review by Alshammari and Bahsoon. This study reviews the state-of-the-art methods for handling imbalanced healthcare data[2], including data sampling techniques, ensemble methods, and deep learning approaches.

Addressing Imbalanced Datasets in Medical Image Analysis by Wang and colleagues. This paper[3] focuses on imbalanced datasets in medical image analysis and proposes a new framework that incorporates both oversampling and undersampling techniques.

Handling Imbalanced Healthcare Data Using Ensemble Methods and Data Sampling Techniques by Al-Bahrani and colleagues. This study[4] proposes an approach for handling imbalanced healthcare data using ensemble methods and data sampling techniques.

A Deep Learning Framework for Handling Imbalanced Medical Data by Wang and colleagues. In this study[6], the authors propose a deep learning framework for handling imbalanced medical data that uses a combination of oversampling and undersampling techniques.

Handling Imbalanced Electronic Health Record Data Using Convolutional Neural Networks with Auxiliary Training by Yao and colleagues. This paper[7] proposes a novel approach for handling imbalanced electronic health record data using convolutional neural networks with auxiliary training.

3. Methodology

All the proposed methodologies applied in previous works were successful in balanced target class and predict Heart Disease or classifying Heart Disease. But, none of the above mentioned methodologies tried to explain how the prediction or classification are made. Why a particular prediction is made in such a way has remained a black box till now. Here, after successfully predicting Heart Disease with the available dataset, we also showed the explanation of the prediction.

3.1 Dataset Description

The dataset[5] contains data from the 2020 annual CDC survey of 400,000 adults regarding their health status. The dataset includes 18 variables, four of which are numerical and 14 of which are categorical. The target class is "Heart Disease," but it is highly imbalanced.

3.2 Data Preprocessing

As earlier mentioned the dataset used here contains both categorical and numerical values. In order to pre-process the data, the proposed model applied label encoding to the categorical features. Using label encoding, it was transform to a numerical value to each class in the categorical features.

3.3 SMOTE

It is a data augmentation technique used in machine learning to balance imbalanced datasets. In many real-world applications, datasets are often imbalanced, which means that the number of instances of one class is significantly higher than the other class. This can lead to biased models, where the model is highly accurate for the majority class but poorly performs for the minority class.

This algorithm takes a point from the minority class, then, it selects it’s nearest neighbors. Then it takes one random point from the neighbors, and calculates the distance between these two points. Now it multiplies the distance with a random fraction between 0 to 1, and generates a new data point.

3.4 Explainable AI

Almost every machine learning classifier(Support Vector Machine, XGBoost, KNN etc.) available today is a black box model, i.e. we don’t really know why we are getting a particular output for a given input value. So, even if we have a good accuracy score in hand, we can’t trust the classifier, there are trust issues.

XAI is one of the most interesting areas of interpreting black box machine learning algorithms. Regression algorithms are most commonly used for interpretability. But complex algorithms give better results, so we need to understand them. LIME and SHAP are the two commonly used algorithms for explaining complex machine learning models.

3.5 LIME(Local Interpretable Model-agnostic Explanations)

LIME provides a local explanation for a particular prediction. It explains which feature has contributed how much and why for a particular row in a dataset. It perturbs the data samples and observes the impact of it on the original data and based on that shows the feature importances of that particular sample.

3.6 SHAP(SHapely Additive exPlanations)

SHAP is a game theory based approach, where it is explained why a particular prediction is different from the baseline and which feature contributed how much to pull or push that prediction value from or to the direction of the base value. So, in a way we can actually debug our Machine Learning model and observe why it predicted a particular prediction. This provides a global explanation.

4. Results and Discussion

Before handling the imbalance data the ratio of majority and minority class i.e. the ratio of 0 and 1 in the target class was 91:9. We have visualized the imbalance in the figure 1.

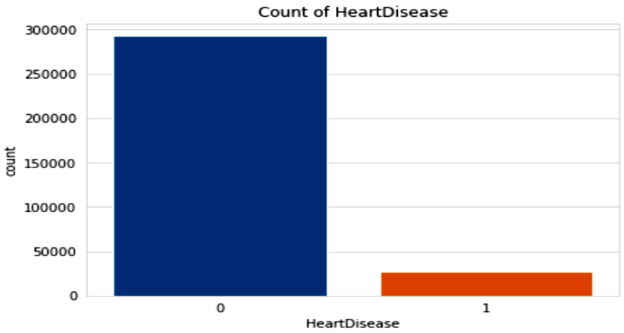


Figure 1: Count plot of heart disease class (Before SMOTE)

First, we applied 4 different algorithms on the initial imbalanced data which are XGBoost, Naive Bayes, Random Forest and KNN. The XGBoost algorithm performed the best with an accuracy of 91.59%(Table 1). But the confusion matrix reveals that it has actually performed poorly on the minority class (Fig 2.). The False Negative rate is 90%, whereas True Negative should be higher. It is clear that it has performed poorly on the minority class i.e. 1.

Table 1: Accuracy Score of different classifiers(Before SMOTE)

Model	Accuracy (%)
XGBoost	**91.59**
Random Forest	**91.10**
KNN	**91.04**
Naive Bayes	**88.06**

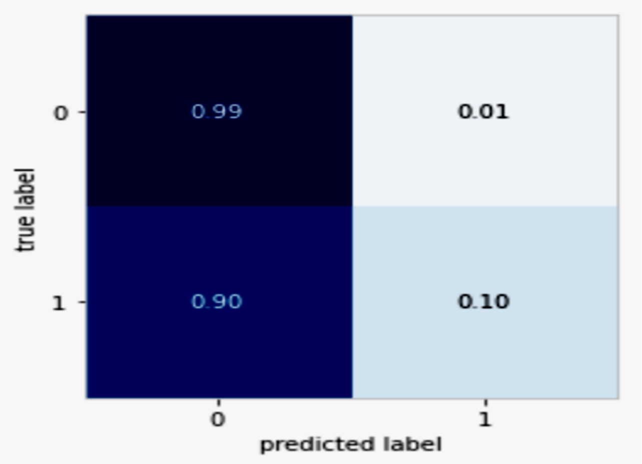


Figure 2: Confusion Matrix of the classification (Before SMOTE)

After applying SMOTE, we can see in fig 3. that the target class imbalance has been taken care of.

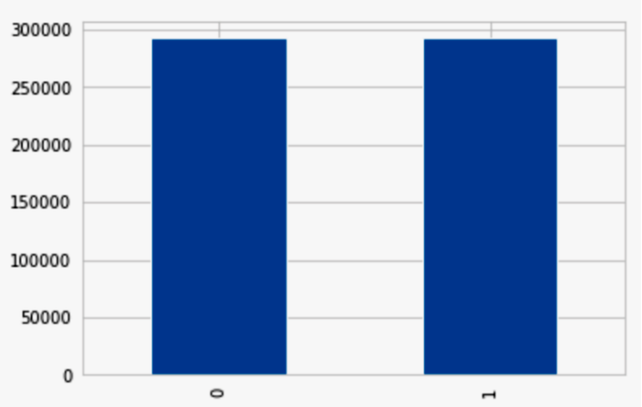


Figure 3: Count plot of heart disease class (After SMOTE)

After handling the imbalance we applied the same 4 algorithms, and gained the results. This time too the XGBoost algorithm performed better than other algorithms with an accuracy of 93.61%(Table 2.). This time the confusion matrix showed more promising results, with a True Positive rate of 99% and a True Negative rate of 91%(fig). We got an AUC value of 98.4%(fig 4.).

Table 2: Accuracy Score of different classifiers (after SMOTE)

Model	Accuracy (%)
XGBoost	**93.61**
Random Forest	**91.39**
KNN	**79.38**
Naive Bayes	**72.63**

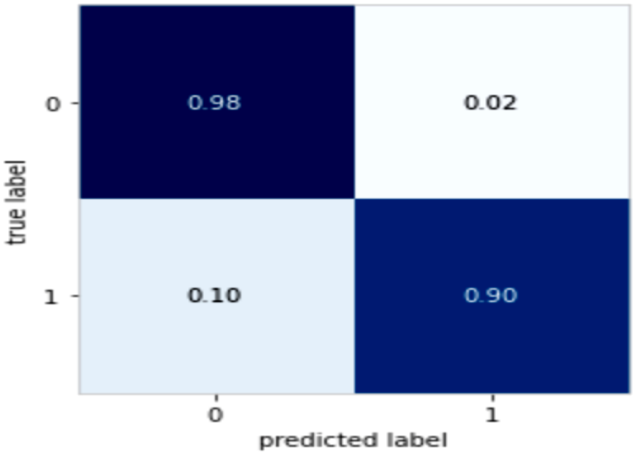


Figure 4: Confusion Matrix of the classification (after SMOTE)

Now, we have applied Explainable AI on the XGBoost model. First we applied SHAP for explaining the global feature importances.

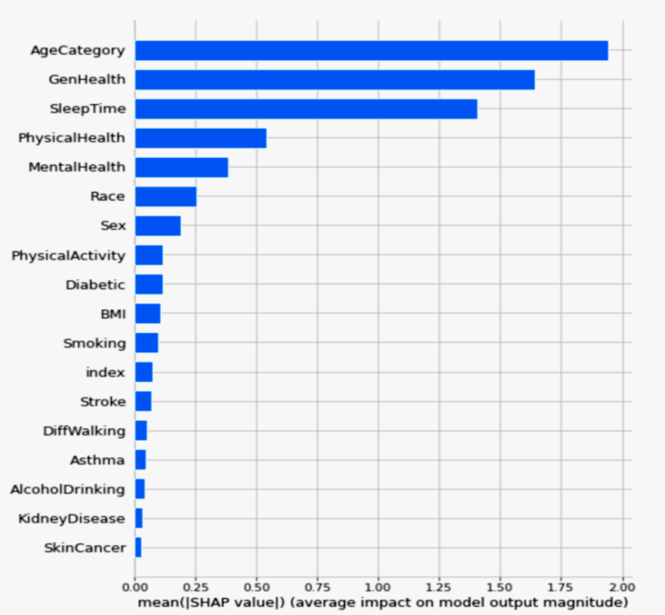


Figure 5: Global feature importance barplot.

Here, in fig 5. we can clearly see that AgeCategory has impacted the most on the prediction globally. And GenHealth, SleepTime, PhysicalHealth and other attributes have provided in a decreasing order.

The force plot for the prediction of the 250th instance(fig 6a.) of data shows why the prediction of a tuple has varied from the base prediction. Here we can see that Sleep time, Age category etc. has pushed the prediction value higher from the base value and Physical health, Physical Activity and others has pushed the prediction value lower from the base value.

Similarly, in fig 6b. we can see that for the 1000th instance age category pushed the base prediction value higher and sleep time, physical health etc. has pushed the value lower.

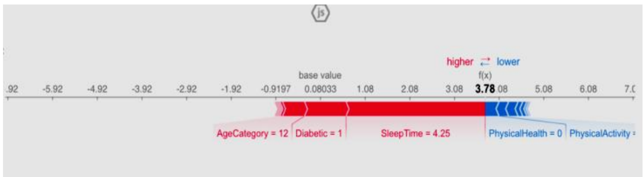


Figure 6a: Shapley values for 250th tuple in the dataset.

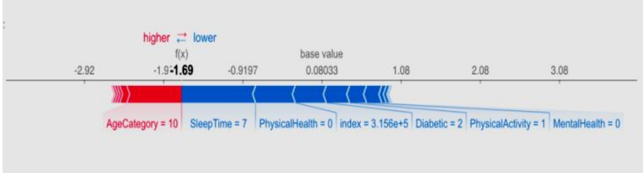
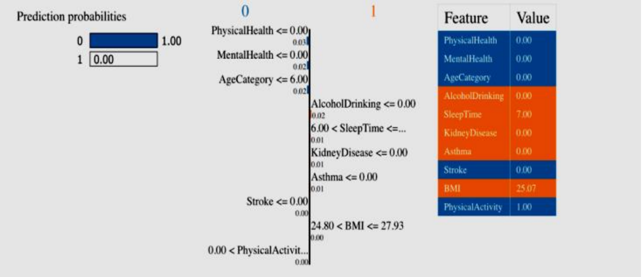


Figure 6b : Shapley values for 1000th tuple in the dataset.

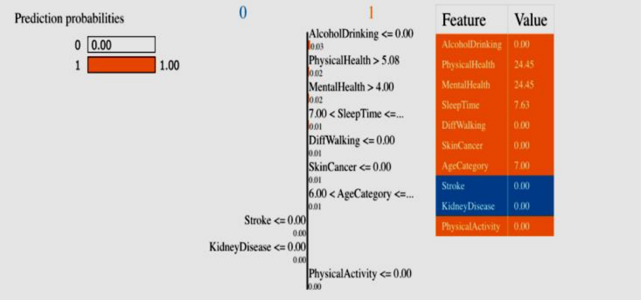
Next, for the local feature importance we have used LIME. In figure 7a. and 7b. we have shown the local feature importance of two tuples. Feature values of each attribute are shown on the right side, using the table.

The blue color in the table means the feature contributed towards the class “0”(no heart disease) and the orange color means the feature contributed towards the class “1”(has heart disease).



Feature	Value
PhysicalHealth	0.00
MentalHealth	0.00
AgeCategory	0.00
AlcoholDrinking	0.00
SleepTime	7.00
KidneyDisease	0.00
Asthma	0.00
Stroke	0.00
BMI	25.07
PhysicalActivity	1.00

Fig 7a: Local explanation for 250th tuple.



Feature	Value
AlcoholDrinking	0.00
PhysicalHealth	24.45
MentalHealth	24.45
SleepTime	7.63
DiffWalking	0.00
SkinCancer	0.00
AgeCategory	7.00
Stroke	0.00
KidneyDisease	0.00
PhysicalActivity	0.00

Fig 7b: Local explanation for 1000th tuple.

For the 250th tuple, “AlcoholDrinking” contributed 0.02 towards the class “1” because it has a value lower than or equal to 0. Other explanations can also be interpreted in the same way.

5. Conclusion and Future Scope

In this work, a highly imbalanced heart disease data is taken for the prediction task. Firstly the “imbalance” has been taken care of, using an oversampling algorithm(SMOTE), where the number of the data in each class is made equal. Additionally, with Explainable AI we have explained the feature importances for the model prediction. According to our work, if the parameters given in the dataset are available for a particular person, we can accurately(93% accuracy) predict if the person has heart disease or not. Though this kind of imbalanced dataset may create a problem for the minority class, it can be taken care of. In future we would try to increase the accuracy of the model and solve the imbalance problem more precisely.

Conflict of Interest

I do not have any conflict of interest.

Authors’ Contributions

Sandip Das[1] is responsible for handling the dataset and training the machine learning model. Gairik Sajjan[2] is tasked with writing the paper and communicating with all the authors. Arkajyoti Poddar[3] will format the paper and assist in writing it. Tamojit Dasgupta[4] will aid in collecting the dataset. Sayani Patty[5] will assist in collecting images. Debmita Ghosh[6] has been supportive throughout the project, including helping to choose the project topic.

Acknowledgements

We take this opportunity to thank Bhabes Bhattacharya, Vice Chancellor of JIS University for giving us the scope of carrying out the project work.

We also express our sincere gratitude to Dr. Dharmpal Singh, Head of the Department of School of Engineering, JIS University and all the respected faculty members of the department of the School of Engineering for giving the scope of successfully carrying out the project work.

References

[1] Deldar, K., Mahdavi, M., & Mohammadzadeh, N. (2020). Handling imbalanced healthcare data with supervised and unsupervised methods: A systematic literature review. Journal of biomedical informatics, 109, 103516.

[2] Alshammari, R., & Bahsoon, R. (2019). Handling imbalanced data in healthcare: A systematic review. ACM Computing Surveys (CSUR), Vol.**52**, Issue.**5**, pp.**1-38, 2019.**

[3] Wang, S., Yao, J., Hu, Y., Zhao, L., & Zhang, Y. (2020). Addressing imbalanced datasets in medical image analysis. IEEE Transactions on Medical Imaging, Vol.**39**, Issue.**7**, pp.**2408-2418, 2020.**

[4] Al-Bahrani, R., Huang, W., & El-Sheimy, N. (2019). imbalanced healthcare data using ensemble methods and data sampling techniques. Applied Sciences, Vol.**9**, Issue.**13**, 2721, **2019.**

[5] https://www.cdc.gov/heartdisease/facts.htm [DATASET]

[6] Wang, H., Yang, X., & Zhang, Q. (2019). A deep learning framework for handling imbalanced medical data. IEEE Access, 7, 89154-89162.

[7] Yao, J., Wang, S., Li, W., & Zhang, Y. (2020). Handling imbalanced electronic health record data using convolutional neural networks with auxiliary training. Journal of biomedical informatics, 110, 103530.

[8] L.H. Yang, J. Liu, Y.M.Wang, L. Martínez, A micro-extended belief rule-based system for big data multiclass classification problems, IEEE Trans. Syst. Man Cybern. Syst. pp.**1–21, 2018.**

[9] P.V. Ngoc, C.V.T. Ngoc, T.V.T. Ngoc, D.N. Duy. A C4. 5 algorithm for english emotional classification, Evolving Syst. 10, pp.**425–451, 2019.**

[10] Datta, Shounak, and Swagatam Das.Near-Bayesian Support Vector Machines forImbalanced Data Classification with Equal or Unequal Misclassification Costs. NeuralNetworks 70: pp.**39–52, 2015.**

[11] ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.114.008729