Effect of Imbalance Data Handling Techniques to Improve the Accuracy of Heart Disease Prediction Using Machine Learning and Deep Learning

Paper ID: 1570789035
Track Name: AI and Blockchain Paradigms for the changing world

Md. Abdus Sahid[1], Mahmudul Hasan[1], Nazrin Akter[2], Md. Motiur Rahman Tareq[1]
[1]Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh
[2]Purdue University, Indiana, USA

Presented by

Md. Abdus Sahid

Presentation Outlines

Introduction

- Approximately *17.9 million* people died of cardiovascular diseases (CVD) in each year according to WHO report, which is **32%** of the fatality rate due to all death reasons. Heart disease is one of the life threatening CVD from the past few years.

- This study focuses on the *effect of different imbalance data handling techniques* to improve the classification accuracy of different **machine learning and deep learning classifiers**.

Motivation and Objective

Nowadays, heart disease is increasing at an alarming rate. Early stage heart disease prediction can help to take proper action to mitigate. The performance of the existing machine learning algorithms are good and different preprocessing techniques are already used in different studies. *Different imbalance data handling techniques have an effect on the classification performance.* To find out the effect, this study balances the dataset using different imbalance data handling techniques and predicts the heart disease. Also, the **best imbalance data handling techniques** for this type of analysis is found out.

01 To perform preprocessing techniques on the heart disease dataset to make it more trainable.

02 To find the effect of different imbalance data handling techniques to the classification performance of machine and deep learning learning classifiers.

03 Comparative study of different machine learning and deep learning classifiers to the classification performance to predict heart disease.

Background Study

Rohit Bharti et al., in 2021 proposed a prediction of heart disease using a combination of machine learning and deep learning [1]. In this study they used a UCI heart disease prediction dataset. They apply Lasso feature selection technique. LR provided 83%, KNN provided 85%, SVM provided 83%, RF provided 80%, DT provided 82% and Deep Learning provided 94% accuracy.

Using feature selection to improve heart disease prediction was proposed by Ke Yuan et al. [2]. They proposed hybrid gradient boosting decision trees with logistic regression (HGBDTLR) ensemble technique. For this study they used Cleveland heart disease dataset and they got accuracy using DT 82%, RF 87%, KNN 64%, AdaBoost 85%, LR 85%, SVM 82%, GBDT 80%, HRFLM 89% and HGBDTLR 92%.

Saiful Islam et al., in 2020 proffered a cardiovascular disease forecast using machine learning paradigms [3]. They used UCI heart disease dataset and applied LR, SVM, DT and NB. Using LR 86%, SVM 84%, DT 75% and NB 74% accuracy they got.

Method and Materials

<u>Cleveland Heart Disease Dataset</u>:

Positive classes - 165(54.46%) and Negative classes - 138(45.54%)

Feature	Description	Feature	Description
age	Age measured in years	thalach	Maximum heart rate achieved
sex	1 = male, 0 = female	exang	Exercise induced angina
cp	Chest pain type: 1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 = asymptomatic	oldpeak	ST depression induced by exercise relative to rest
trestbps	Resting blood pressure (in mm Hg on admission to the hospital)	slope	The slope of the peak exercise ST segment 1 = upsloping 2 = flat 3 = down sloping
chol	Serum cholesterol in mg/dl	ca	Number of major vessels (0-3) colored by fluoroscopy
fbs	Fasting blood sugar > 120 mg/dl 1 = true; 0 = false	thal	3 = normal 6 = fixed defect 7 = reversible defect
restecg	0 = normal 1 = having ST-T wave abnormality 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria	Class	Diagnosis of heart disease (angiographic disease status) 0 = < 50% diameter narrowing 1 = > 50% diameter narrowing

Method and Materials

Imbalance Data Handling Technique

We employed several imbalance data handling techniques. These are listed below:

1. SMOTE
2. ADASYN
3. SMOTETomek
4. NearMiss

Classifier

For this study we employed six classifiers. These are listed below:

1. Support Vector Machine (SVM)
2. Gaussian Naive Bayes (GNB)
3. Random Forest (RF)
4. Logistic Regression (LR)
5. Multilayer Perceptron (MLP)
6. LR-MLP Ensemble

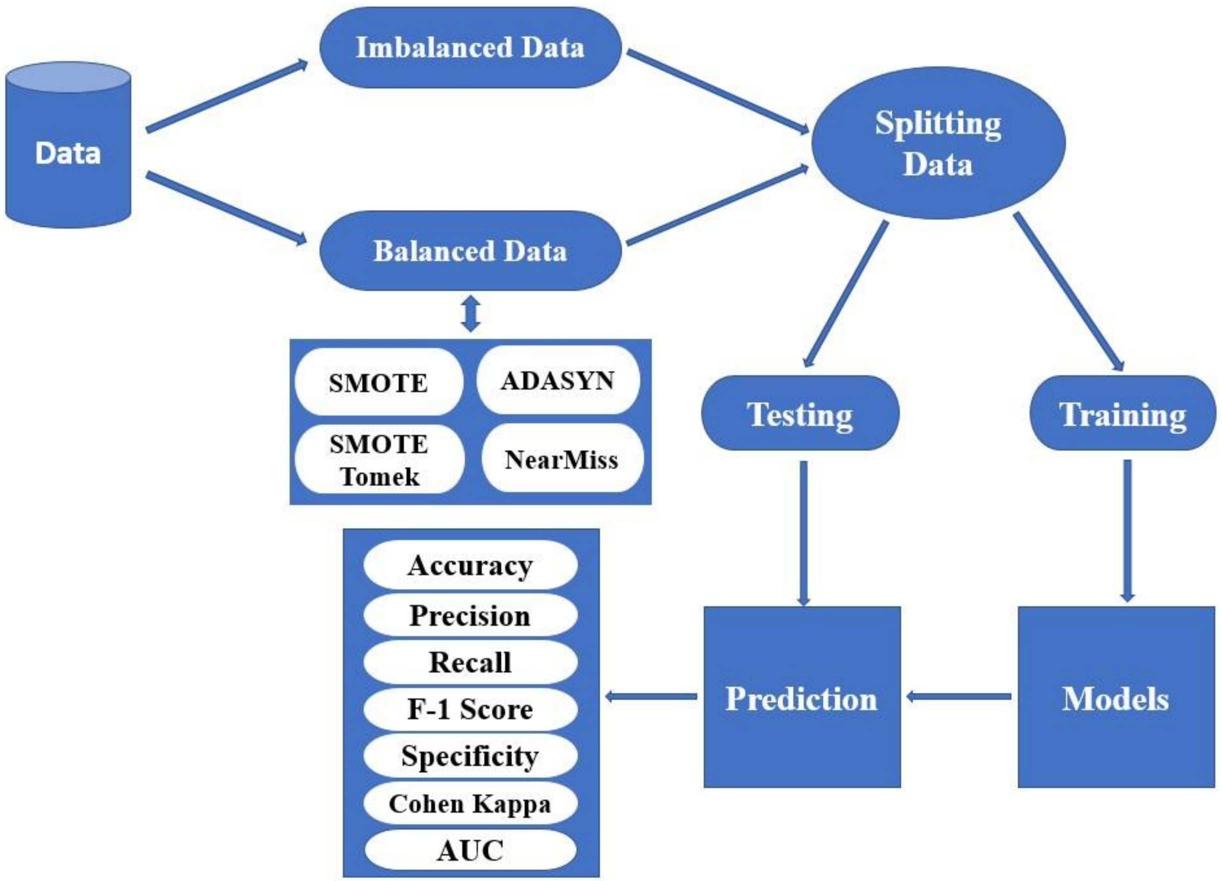
Evaluation Metric

Several evaluation metric we used to measure the performance. These are listed below:

1. Accuracy
2. Precision
3. Recall
4. F-1 Score
5. Specificity
6. Cohen Kappa
7. AUC Score
8. ROC Curve

Method and Materials

Methodology



Step 1

Data Collecting

Step 2

Data Processing

Step 3

Imbalance data handle

Step 4

Training and Testing

Step 5

Performance Evaluation

Step 6

Result and discussion

Result Analysis

Performance on Imbalance data

	Algorithm	Accuracy	Preci-sion	Recall	F-1 Score	Specifi-city	Cohen Kappa	AUC
Imbalanced Dataset	SVM	95%	0.95	0.95	0.95	0.92	0.9	0.96
	LR-MLP	92%	0.92	0.92	0.92	0.88	0.83	0.96
	LR	92%	0.92	0.92	0.92	0.88	0.83	0.96
	GNB	90%	0.9	0.9	0.9	0.88	0.8	0.94
	MLP	90%	0.9	0.9	0.9	0.92	0.8	0.96
	RF	89%	0.89	0.89	0.89	0.84	0.76	0.94

Performance on balanced data using

S

	Algorithm	Accuracy	Preci-sion	Recall	F-1 Score	Specifi-city	Cohen Kappa	AUC
SMOTETomek Balanced Dataset	SVM	96%	0.96	0.96	0.96	0.94	0.91	0.97
	LR-MLP	96%	0.96	0.96	0.96	0.97	0.91	0.96
	LR	94%	0.94	0.94	0.94	0.97	0.88	0.96
	GNB	94%	0.94	0.94	0.94	0.97	0.88	0.95
	MLP	96%	0.96	0.96	0.96	0.97	0.91	0.97
	RF	94%	0.94	0.94	0.94	0.97	0.88	0.97

Performance on balanced data using SMOTE

	Algorithm	Accuracy	Preci-sion	Recall	F-1 Score	Specifi-city	Cohen Kappa	AUC
SMOTE Balanced Dataset	SVM	94%	0.94	0.94	0.94	0.97	0.89	0.97
	LR-MLP	92%	0.92	0.92	0.92	0.92	0.83	0.96
	LR	94%	0.94	0.94	0.94	0.94	0.89	0.96
	GNB	93%	0.93	0.93	0.93	0.94	0.86	0.95
	MLP	92%	0.92	0.92	0.92	0.92	0.83	0.96
	RF	93%	0.93	0.93	0.93	0.94	0.86	0.95

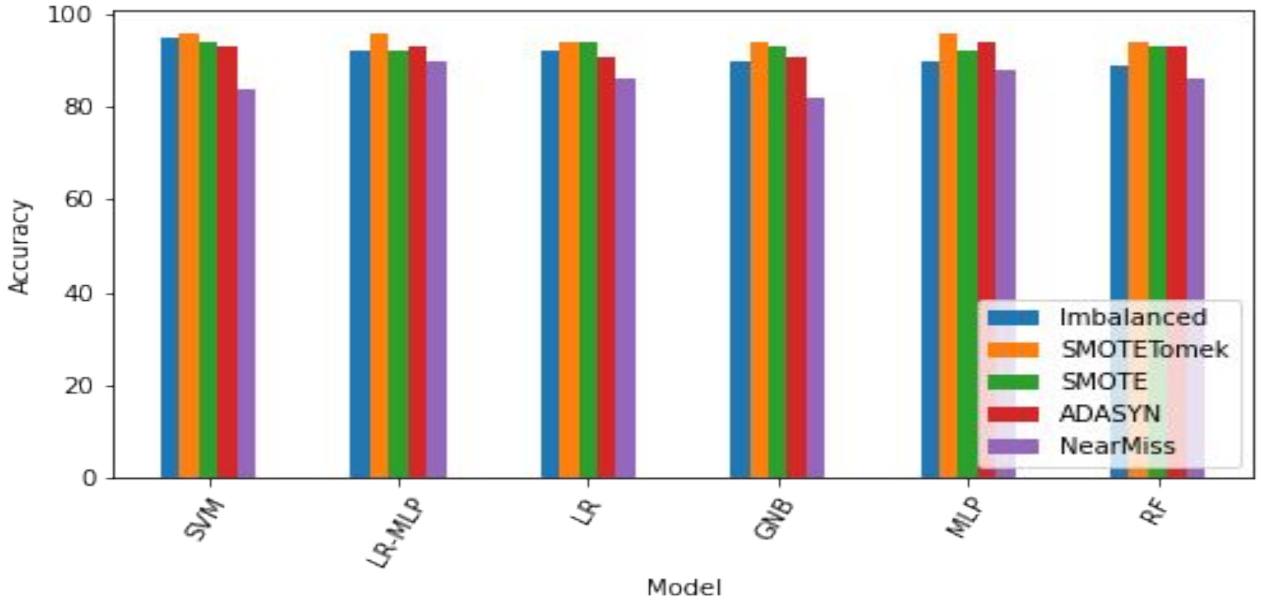
Performance on balanced data using ADASYN

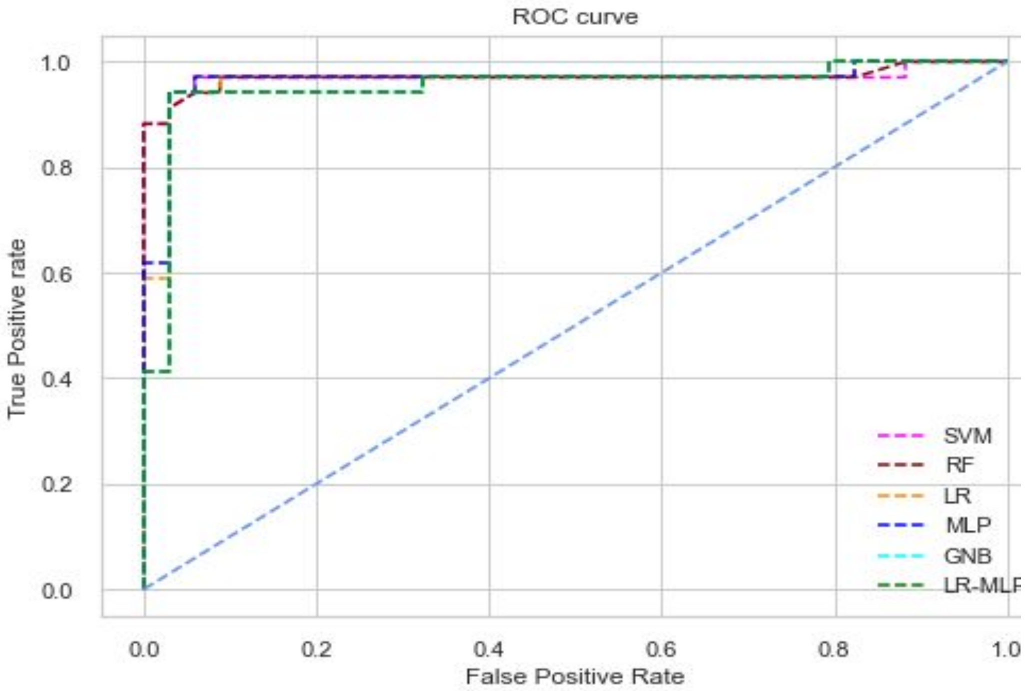
	Algorithm	Accuracy	Preci-sion	Recall	F-1 Score	Specifi-city	Cohen Kappa	AUC
ADASYN Balanced Dataset	SVM	93%	0.93	0.93	0.93	0.94	0.85	0.97
	LR-MLP	93%	0.93	0.93	0.93	0.94	0.85	0.96
	LR	91%	0.91	0.91	0.91	0.91	0.83	0.96
	GNB	91%	0.91	0.91	0.91	0.94	0.83	0.94
	MLP	94%	0.94	0.94	0.94	0.97	0.88	0.96
	RF	93%	0.93	0.93	0.93	0.94	0.85	0.95

Result Analysis

NearMiss Balanced Dataset	Algorithm	Accuracy	Preci-sion	Recall	F-1 Score	Specifi-city	Cohen Kappa	AUC
	SVM	84%	0.85	0.84	0.84	0.92	0.68	0.9
	LR-MLP	90%	0.9	0.9	0.9	0.92	0.8	0.94
	LR	86%	0.86	0.86	0.86	0.88	0.72	0.93
	GNB	82%	0.8	0.8	0.8	0.84	0.6	0.89
	MLP	88%	0.88	0.88	0.88	0.92	0.76	0.92
	RF	86%	0.87	0.86	0.86	0.92	0.72	0.91

Result are shown by the bar chart and ROC curve.





Conclusion and Future Works

CONCLUSION

- This analysis mainly focuses on the *effect of different imbalance data handling techniques* to improve the accuracy to predict the heart disease using Cleveland dataset. Most of the cases SVM performs well than other models.
- In imbalance data most of the algorithms shows greater than or equal to **89% accuracy.** But, using **SMOTETomek** imbalance data handling techniques, all the algorithms *shows* greater than or equal to *94% accuracy.*
- Also performance of SMOTE (Oversampling), ADASYN (Oversampling) is better than imbalance data but the performance of NearMiss (Under sampling) is too low compared to other techniques.

FUTURE WORKS

- Feature selection and dimensionality reduction may improve the analysis in this sector.
- Hybrid and Ensemble technique may also improve the accuracy level of prediction.
- Deep learning algorithms plays a vital role in the field of healthcare. Other deep learning algorithms may give better outcome.

References

1. Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S. and Singh, P., "Prediction of Heart Disease using a Combination of Machine Learning and Deep Learning," Computational intelligence and neuroscience, 2021, doi.org/10.1155/2021/8387680.

2. Uyar, K. and İlhan, A., "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks," Procedia Computer Science, 2017, Vol. 120, pp.588-593.

3. S. Islam, N. Jahan and M. E. Khatun, "Cardiovascular Disease Forecast using Machine Learning Paradigms," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC),2020, pp.487- 490,doi:10.1109/ICCMC48092.2020.ICCMC-00091

Thank You

Question & Answer