Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 6, 2021, Pages. 20340 – 20347
Received 20 August 2021; Accepted 10 September 2021.

An Improved Framework of Liver Disease Detection using SMOTE + ENN

N Shubhankar[1*], Mayank Gupta[2], M Gayathri[3]
[1, 2, 3] SRMIST, Kattankulathur, India
[*] shubhankarnandakumar07@gmail.com

ABSTRACT

The diagnosis of liver disease at a preliminary stage is consequential for convalescent treatment. It is a demanding task for medical researchers to predict the early stages' condition due to indistinct symptoms. More often than not, the symptoms become evident when it is too late. Furthermore, with the rise in the number of tests, the demands for faster and more accurate test systems have also risen. To overcome this issue, machine learning classification algorithms have been implemented on various datasets of liver patients. However, these datasets have a disproportionate number of cases for each class, making the model biased. We aim to solve this imbalance by applying a hybrid approach of SMOTE oversampling and Edited Nearest Neighbour undersampling techniques, which provide much cleaner data due to aggressive undersampling by ENN.

KEYWORDS

Classification Algorithm, Liver diseases, Machine Learning, Imbalance dataset, SMOTE-ENN.

Introduction

The liver is an essential organ of human anatomy. It is responsible for regulations of chemical levels in the blood. Liver diseases are prevalent and can occur due to various activities such as drinking alcohol. With the rise in patient numbers, a need has arisen for faster diagnosis, which has seen researchers venture into AI and ML. However, most of the present medical datasets have a significant imbalance wherein one class's quantity far exceeds the other. In such cases, the model tends to perform poorly for the infrequent type. This becomes rather compound in binary classification matters wherein the classification paradigm showcases higher classification error of the minority class than the majority class.

Different techniques help us solve the class imbalance by generating minority class cases from the existing dataset and balance the number of points between the majority class and the minority class. Different techniques tackle this in different ways and affect the overall dataset differently. In this dataset, we use a hybrid approach of SMOTE(Synthetic Minority Oversampling Technique) with Edited Nearest Neighbor (ENN) on Indian Liver Patient Dataset(ILPD). For medical diagnosis, classification algorithms are generally preferred, and in this paper, we have used four algorithms, namely Logistic Regression, Random Forest, Multilayer Perceptron, and Support Vector Machines.

Literature Review

Michael J Zorich et al. [1] suggested in their paper that Support Vector Machines (SVM)[19] classifiers produced the best result in cases of chemical datasets. Lung-Cheng Huang et al. [2] suggested that the Naive Bayes Classifier produced a higher performance than Support Vector Machines and C4.5 Algorithms for Chronic Fatigue Syndrome Dataset (CFD). However, Paul Harper et al. [3] reported that instead of there being a single best classifier, the title depended on the dataset's features to be analyzed. Joel Jacob et al. [4] aimed to compare classification algorithms based on their performance factors. They have used Support Vector Machine, Logistic Regression, K - Nearest Neighbour (KNN), and Artificial Neural Network (ANN). ANN performed the best with 92.8% accuracy.

P. Kuppan et al. [5] analyzed liver disorder data incorporating case histories such as J48, Naive Bayes, Decision Table. They drew conclusions such as men were more likely to suffer from the disease than women, age group 35-

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 6, 2021, Pages. 20340 – 20347
Received 20 August 2021; Accepted 10 September 2021.

65 had the highest density of cases, and smoking and drinking activities resulted in diseases 24% and 26%, respectively.

Pushpendra Kumar et al.[6] tried to counter the imbalance problem by using SMOTE oversampling. They used two datasets: The Indian Liver Patient Dataset (ILPD) and Madhya Pradesh Region Liver Patient Dataset (MPRLPD). When applied to these datasets before being fed to the machine learning algorithms, SMOTE improved accuracy, specificity, and sensitivity for both algorithms. SVM had the highest accuracy for both datasets, 96.42% for MPRLPD and 73.96% for ILPD. However, the application of SMOTE showed a sharp decline in precision values. They tried a new approach [7] using variable- neighbor weighted fuzzy K nearest neighbor approach and Tomek Link-Redundancy Undersampling (TLRUS), which showed better results than most existing papers.

Proposed Work

Logistic Regression

It is one of the simplest classification models. The Logistic Function [8] is used to look at the relationship between various parameter variables. If the parameters are known, it is possible to recreate the entire model. It follows the equation;-

$$\text{output} = e^{(b + c*x)} / (1+e^{(b+ c*x)})$$

Where b is the bias or intercept, and c is the coefficient.

Random Forest

It is an ensemble learning method implemented using a large number of decision trees. Introduced in the year 1995[9], it overcomes decision trees' limitations to overfit data and thus performs much better. When working on classification, Random forest generally tends to work on Entropy or Gini Index Values whose formula is given below.

$$\text{Gini Index} = 1 - \sum_{1}^{c} (p_i)^2$$

Where c is the number of classes and pi is the relative frequency.

$$\text{Entropy} = \sum_{1}^{c} - p_i * \log_2(p_i)$$

Where c is the number of classes and pi is the relative frequency.

Multilayer Perceptron

It is a class of feedforward neural networks that are used to distinguish non-linearly separable data. It generally consists of three layers, namely input, hidden, and output. Each layer has a set of neurons that has their activation function. These neurons are trained by a method called backpropagation. The activation function used is Rectified Linear Units (ReLU).

$$f(x) = x^{+} + \max(0,x)$$

where x is the input.

Support Vector Machine

It is a supervised model introduced in 1997 by Vapnik et al. [10] that uses a hyperplane to maximize margins between the classes. It uses a kernel function to provide a higher dimension in the case of non-linear variables. The optimal hyperplane is calculated using,

$$W.Xi + b = 0$$

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 6, 2021, Pages. 20340 – 20347
Received 20 August 2021; Accepted 10 September 2021.

Wherein W is the weight vector for individual tuple Xi and bias b. The sides of the margin can be defined by adjusting the weights using.

$$yi(W.Xi + b) \geq 1 \ \forall i$$

The hyperplane can then be maximized using

$$\text{Min } \|W\| \text{ or Max} \frac{2}{\|W\|}$$

SMOTE (Synthetic Minority Oversampling Technique)

It is the most common method used to oversample the minority class to counter the class imbalance. This method was first put forth by Nitesh Chawla et al.[11] in 2005. This oversampling is done by finding feature space similarities of neighbors and synthesizing samples between them. It can be represented using the equation,

$$X_{sym} = X_i + (\hat{X}_i - X_i) * \delta$$

Where Xi is a minority instance, and $\hat{X}$ is a random neighbor of its.

Edited Nearest Neighbour

It is a standard method used to undersample the majority class to counter the class imbalance. It was first implemented by Dennis Wilson [12] in 1972. This is an aggressive form of undersampling, as mentioned by Gustavo Batista et al. [13], wherein ENN can be used to remove samples from both majority and minority classes. This cleansing is done using any three nearest neighbors in the dataset to find samples that have been misclassified.

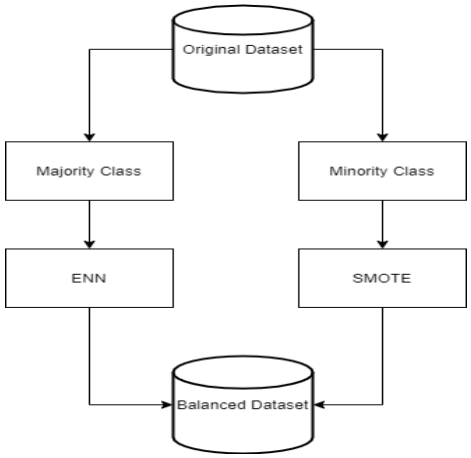


Figure 1. Hybrid Approach in balancing the dataset.

Dataset

The dataset intended to be used is the Indian Liver Patient Dataset (ILPD) from the UCI ML repository [14] collected from North East Andhra Pradesh. It contains 583 entries on 10 different attributes. There are 416 entries with the patient suffering from liver disease and 167 patients free from liver disease. Table 2 shows the type of each attribute.

Bilirubin, Albumin, Alkaline Phosphatase, Alanine Aminotransferase, Aspartate Aminotransferase, and globulin are the various chemicals present in the liver whose varying levels can be used to identify or diagnose liver diseases. The final result column is binary data, with 1 indicating presence and 2 indicating the absence of illness.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 6, 2021, Pages. 20340 – 20347
Received 20 August 2021; Accepted 10 September 2021.

Table 1. Attributes In Dataset

SNO	ATTRIBUTE	TYPE
1	Age	Numeric
2	Sex	Nominal
3	Total Bilirubin	Numeric
4	Direct Bilirubin	Numeric
5	Alkaline Phosphatase	Numeric
6	Alanine Aminotransferase	Numeric
7	Aspartate Aminotransferase	Numeric
8	Total Proteins	Numeric
9	Albumin	Numeric
10	Albumin and Globulin Ratio	Numeric
11	Result	Numeric(1,2)

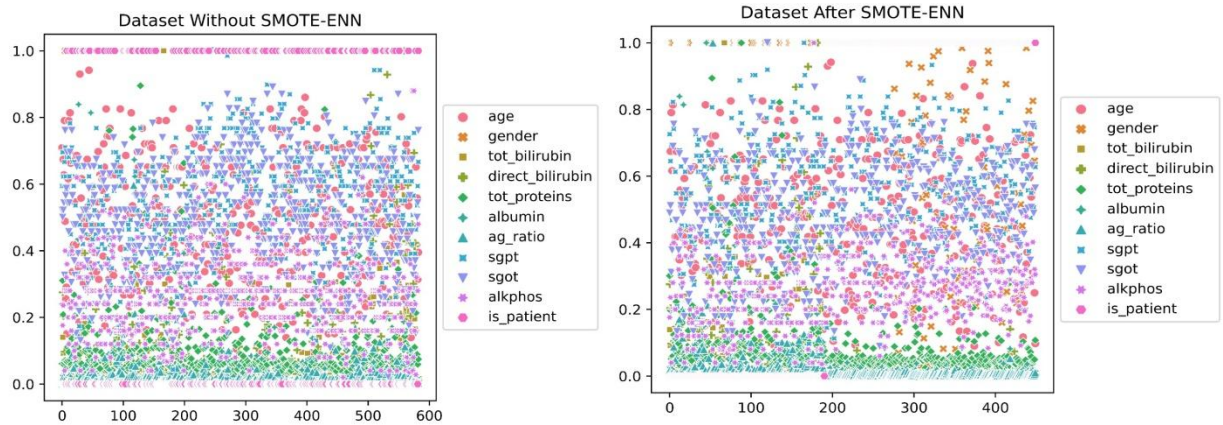


Figure 2. (a)Original Dataset (b)Dataset with Application of SMOTE+ENN

Evaluation Parameters

These metrics are instrumental in evaluating the performance of the models. Three metrics have been used for each algorithm, namely Precision, Sensitivity or Recall, and AUC-ROC.

- Precision: This is an evaluation of the count of True Positive among all positive values.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- Sensitivity or Recall: This is an evaluation of the count of True Positive among all values classified correctly.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

- AUC-ROC: It stands for Area Under Curve (AUC) and Receiver Operating Characteristics (ROC). It plays a fundamental role in evaluating the performance of classifiers.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 6, 2021, Pages. 20340 – 20347
Received 20 August 2021; Accepted 10 September 2021.

Implementation

The four algorithms - Logistic Regression, Multilayer Perceptron, Random Forest, and SVM were trained on the dataset with and without the hybrid SMOTE+ENN and logged their metrics.

- Load the dataset from the UCI repository and pre-process the dataset to ensure no NAN values are present and convert gender from nominal to numeric.
- Apply the SMOTE+ENN hybrid method on the dataset to balance it.
- Train all the models on all four algorithms using both datasets.
- Log the average for all metrics for each of the models.

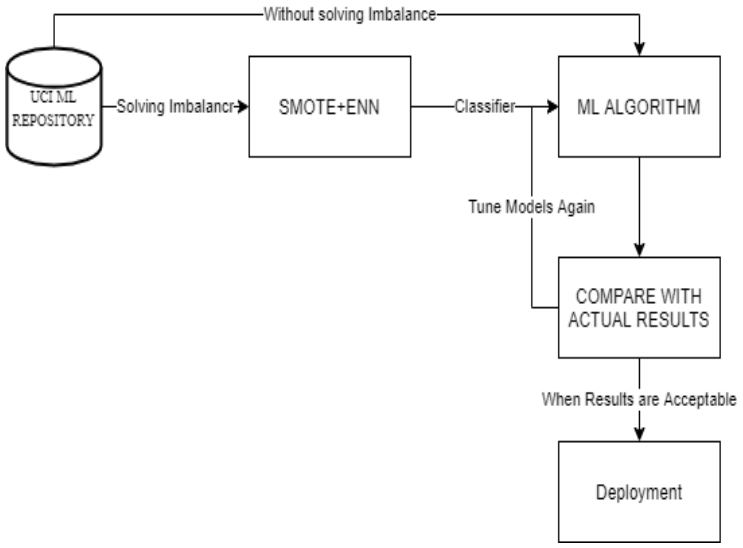


Figure 3. Architecture Diagram of Implementation

Results

The logged metrics for all four algorithms can be seen in Table 2, Table 3, Table 4, and Table 5.

Table 2. Logistic Regression

METRICS	LOGISTIC REGRESSION	
	Imbalanced Dataset	**Balanced Dataset**
Precision	0.746	0.863
Sensitivity	0.944	.0504
AUC-ROC	0.755	0.730

According to Table II, Logistic Regression gave better results in Precision 0.863 with a balanced dataset, whereas it gave better results in Sensitivity 0.944 and AUC-ROC 0.755 with the imbalanced dataset.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 6, 2021, Pages. 20340 – 20347
Received 20 August 2021; Accepted 10 September 2021.

Table 3. Random Forest

METRICS	RANDOM FOREST	
	Imbalanced Dataset	Balanced Dataset
Precision	0.735	0.906
Sensitivity	1.000	0.624
AUC-ROC	0.768	0.792

According to Table III, Random Forest gave better results in Precision 0.906 and AUC-ROC 0.792 with a balanced dataset, whereas Sensitivity 1.000 was much better with the imbalanced dataset.

Table 4. MLP

METRICS	MLP	
	Imbalanced Dataset	Balanced Dataset
Precision	0.766	0.875
Sensitivity	0.760	0.560
AUC-ROC	0.704	0.723

According to Table IV, Multilayer Perceptron performed better in the parameters Precision 0.875 and AUC-ROC 0.723 with Balanced dataset, whereas it performed better in Sensitivity 0.760 with the imbalanced dataset.

Table 5. SVM

METRICS	SVM	
	Imbalanced Dataset	Balanced Dataset
Precision	0.725	0.859
Sensitivity	0.824	0.440
AUC-ROC	0.652	0.723

According to Table V, Support Vector Machine performed better in Precision 0.859 and AUC-ROC 0.723 with a balanced dataset, whereas it performed better in Sensitivity 0.824 with Imbalanced dataset.

From the below charts, a number of conclusions can be made. Precision values for all four models have gone up on the balanced dataset indicating that the count of false-positive or FP cases has gone down. Similarly, for all models except Logistic Regression, AUC-ROC values have gone up on the balanced dataset. However, for all four models, the value of Sensitivity has gone down on the balanced dataset, indicating a rise in false-negative or FN cases.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 6, 2021, Pages. 20340 – 20347
Received 20 August 2021; Accepted 10 September 2021.

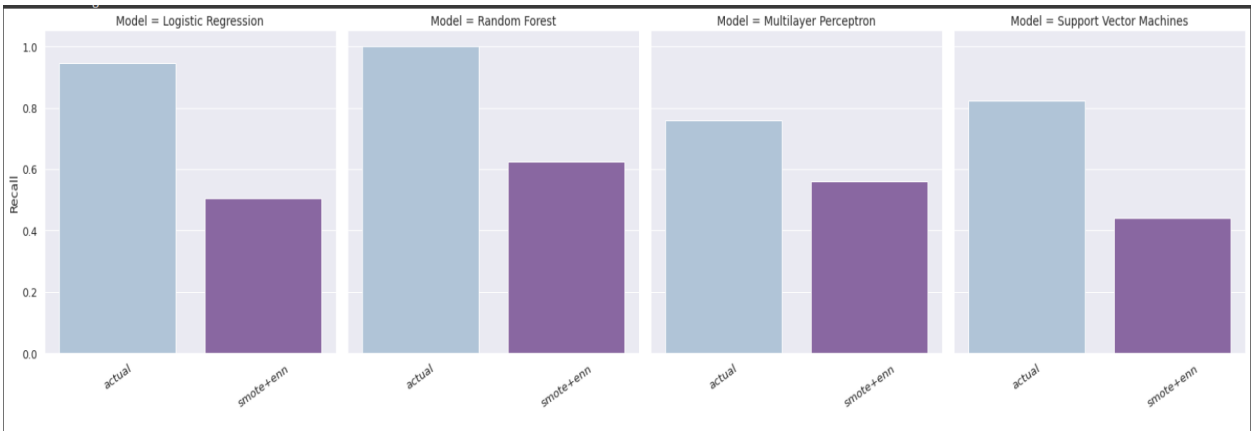


Figure 4. It shows a comparison bar plot of Sensitivity values for all four models with and without the application of SMOTE+ENN

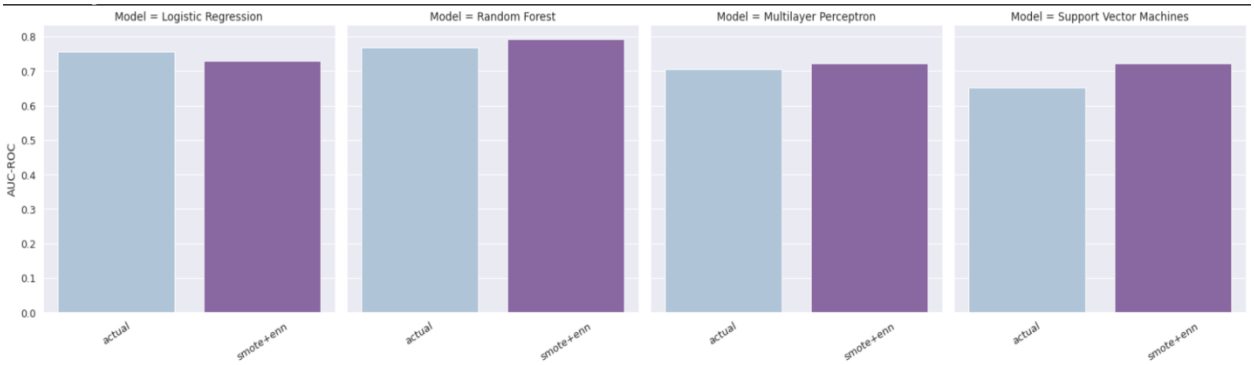
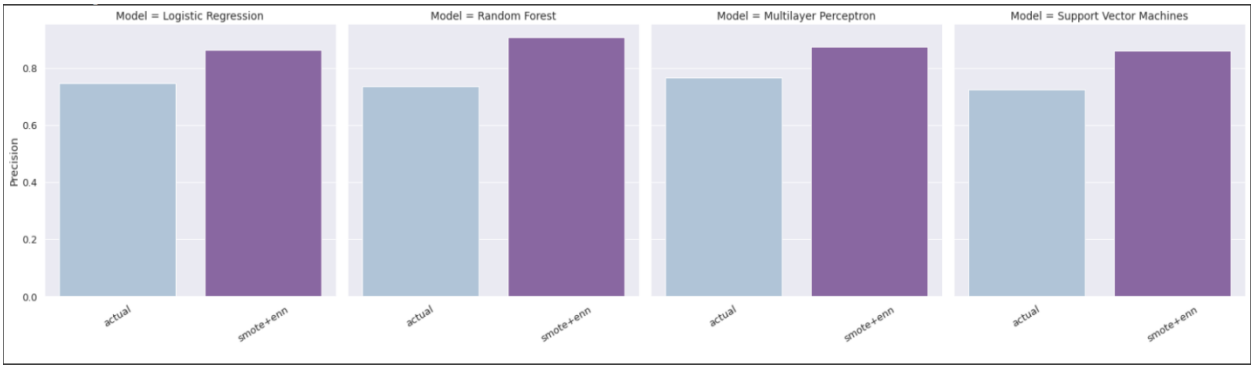


Figure 5. It shows a comparison bar plot of AUC-ROC values for all four models with and without the application of SMOTE+ENN



.**Figure 7.** It shows a comparison bar plot of Precision values for all four models with and without the application of SMOTE+ENN

Conclusion

This paper has applied SMOTE and ENN's hybrid approach to balance the dataset to diagnose the liver disorder. Imbalanced datasets tend to become biased, and that is unfavorable for such situations. The proposed system has shown some improvement from existing models and makes a case for this Hybrid approach. Future works may include improving the model with the use of gradient boosting algorithms and other intelligent algorithms.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 6, 2021, Pages. 20340 – 20347
Received 20 August 2021; Accepted 10 September 2021.

References

[1] Michael J. Sorich, John O. Miners, Ross A. McKinnon, David A. Winkler, Frank R. Burden, and Paul A. Smith, “Comparison of linear and nonlinear classification algorithms for the prediction of drug and chemical metabolism by human UDP- Glucuronosyltransferase Isoforms”

[2] Lung-Cheng Huang, Sen- Yen Hsu and Eugene Lin, “A comparison of classification methods for predicting Chronic Fatigue Syndrome based on genetic data”, (2009)

[3] Paul R. Harper, “A review and comparison of classification algorithms for medical decision making.”

[4] Joel Jacob, Joseph Chakkalakal Mathew, Johns Mathew, Elizabeth Issac, “Diagnosis of Liver Disease Using Machine Learning Techniques”, International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 04, Apr-2018

[5] P. Kuppan, N. Manoharan, “A Tentative analysis of Liver Disorder using Data Mining Algorithms J48, Decision Table and Naive Bayes”, International Journal of Computing
Algorithm, 6 (1) (2017), pp. 2239-2278

[6] Pushpendra Kumar, Ramjeevan Singh Thakur, “ Early Detection of the Liver Disorder from Imbalance Liver Function Test Datasets”, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-4, February 2019

[7] Liver disorder detection using variable- neighbor weighted fuzzy K nearest neighbor approach Pushpendra Kumar, Ramjeevan Singh Thakur - Springer Science+Business Media, LLC, part of Springer Nature 2020

[8] Cramer, J. S. (2002). The origins of logistic regression (PDF) (Technical report). 119. Tinbergen Institute. pp. 167–178.

[9] Ho, Tin Kam (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

[10] Vapnik, Vladimir N. (1995). "Support-vector networks" (PDF). Machine Learning. 20 (3): 273–297.

[11] SMOTE: synthetic minority over-sampling technique - Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer Journal of Artificial Intelligence ResearchJune 2002

[12] Asymptotic Properties of Nearest Neighbor Rules Using Edited Data Dennis L. Wilson, IEEE Transactions on Systems, Man, and Cybernetics (Volume: SMC-2, Issue: 3, July 1972)

[13] A study of the behavior of several methods for balancing machine learning training data, Gustavo E. A. P. A. Batista,Ronaldo C. Prati, Maria Carolina Monard - ACM SIGKDD Explorations NewsletterJune 2004

[14] https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)