Section: Research Paper

A Hybrid SMOTE-NMA Model for Predicting Stage of Liver Disease

1.K.Sindhya Research Scholar, Department of Computer Science,
Rathnavel Subramaniam college of Arts and Science, Sulur, Coimbatore, Tamilnadu, India.
2.Dr.M.Suganya Associate Professor and Head, Department of Information Technology.
Rathnavel Subramaniam college of Arts and Science, Sulur, Coimbatore, Tamilnadu, India.

Abstract: - Liver illnesses, especially liver cirrhosis, acute liver failure, and liver cancer, account for about 2 million annual fatalities globally. India experiences problems with the quality of the epidemiological data resources for liver illness in terms of diagnostic precision and clinical phenotyping, uniformity of reporting, and the absence of national electronic databases, much like many other developing nations. Liver failure can be prevented by detecting and treating liver issues early on. With a data imbalance, it is essential to predict the stage of liver disease. The model's novelty is its use of the hybrid SMOTE-NMA model to manage imbalanced data for both majority and minority classes. Precision, recall, f1-score, fitting time, and test accuracy are used to gauge how well the models work. With the base models LR, RC, and GNB, the proposed model HSMOTENMA had the highest recall of 0.88, 0.94, and 0.82.

Keyword:- Liver Disease, Imbalanced Data, SMOTE, NMA, Liver Disorder Stage Prediction

1. Introduction

The largest organ in the body, the liver (Jigger), functions like an engine and carries out more than 500 essential tasks, including the production of bile and several proteins, the storage of glycogen as a ready source of blood glucose, the metabolism of nutrients, the detoxification of drugs and other noxious substances like NH3, and the control of blood clotting. Hepatitis viruses (A through E), alcohol, non-alcoholic fatty liver disease (NAFLD), medications, autoimmune and genetic illnesses, cryptogenic, and benign and malignant liver tumours are the main causes of liver diseases. Over 2 million deaths worldwide occur each year as a result of liver diseases, primarily liver cirrhosis, acute liver failure, and liver cancer.

In India, one in every five persons has liver disease, which is rapidly spreading like an epidemic. Incredibly, India now accounts for 268,580 (3.17% of all deaths) liver-related deaths annually, making up 18.3% of the 2 million liver-related deaths worldwide. India is quickly realizing that liver ailments are a public health priority. Similar to many other developing countries, India has issues with the quality of the epidemiological data resources for liver disease in terms of diagnostic accuracy and clinical phenotyping, uniformity of reporting, and the lack of national electronic databases [6]. By identifying and treating liver problems at an early stage, liver failure can be avoided. There are four phases of liver disease, the first of which is characterized by inflammation and may or may not be accompanied by any symptoms in the patient. Long-term inflammation leads the illness to progress into the second stage, called fibrosis, which is also mostly asymptomatic, by replacing the good liver tissue with scar tissue. Cirrhosis, which is the third stage, is brought on by severe liver scarring. At this point, the patient begins to exhibit symptoms including nausea, vomiting, weakness, jaundice, and stomach discomfort. End-stage

liver disease is present when there has been a significant decline in liver function (ESLD). In this case, the patient exhibits serious problems, yet they are treatable without a liver transplant. Because it is the most important and major organ in the body, maintaining its health is key for better overall health. But the truth is that when it comes to health, individuals frequently ignore it [11]. The majority of people worldwide have mild to severe liver disorders as a result of bad lifestyle choices. Since the epidemiological statistics are not as reliable as they are in most other regions of the world, it is challenging to determine the prevalence of liver disease in India. However, the information that is currently available indicates that the mortality rate in India is rising due to cirrhosis and its consequences. The leading causes of cirrhosis and liver cancer-related mortality are probably hepatitis B and C, alcoholic liver disease, and non-alcoholic liver disease. India has a number of issues that contribute to the spread of disease, such as a lack of resources and healthcare facilities, cultural beliefs, a reliance on unproven traditional practices and herbal remedies, a lack of knowledge about the transmission of infections that cause liver diseases, and unfavorable socioeconomic conditions.

1.1.Stages of Liver Disease

The liver is a remarkable organ. Liver can heal and even rebuild itself if it is diagnosed when some scar tissue has already started to form. Because of this, liver disease damage can frequently be repaired with an effective treatment strategy. The four phases of liver disease (as shown in fig. 1.1), often known as end-stage liver disease, start with inflammation and develop all the way to liver failure.

- Stage 1: Inflammation- Liver will be inflamed and sometimes painful in the early stages.
- Stage 2: Scarring and Fibrosis- Scarring will result from the inflammation if it is not treated. Liver's blood flow is restricted as scar tissue accumulates, which prevents the good sections from performing their functions and forces them to work harder.

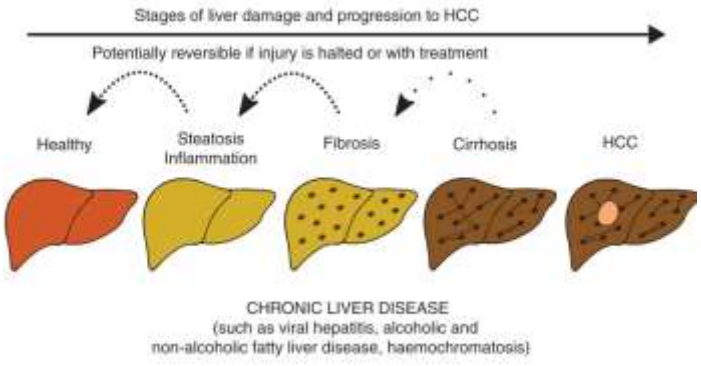

Figure: 1.1. Stages of Liver Disease (Source: https://www.ncbi.nlm.nih.gov/books/NBK549197)

- Stage 3: Cirrhosis- Liver will either not function at all or will function poorly as the scar tissue takes over and leaves less and less healthy tissue to perform its functions.
- Stage 4: Advanced liver disease or failure- This is a catch-all word for a number of ailments, including swollen liver, internal bleeding, kidney function loss, fluid retention, and lung issues.

1.2.Motivation

Every year, millions of people die from liver illnesses. Each year, viral hepatitis alone results in 1.34 million fatalities. As the liver will continue to operate correctly even when it is partially damaged, problems with liver patients are difficult to identify at an early stage. Indians are at a

significant risk of liver failure. By 2025, it's possible that India may become the world's epicenter for liver diseases. Due to a desk-bound lifestyle, increasing alcohol intake, and smoking, liver infections are very common in India. Patients' chances of survival will rise with an early liver stage diagnosis. Given these concerning numbers, it is important to be concerned about combating these diseases. After all, having ill adolescents is not feasible for a developed and affluent country.

1.3.Contribution

Predicting liver disease stage is crucial with an imbalanced data. In other words, if dataset includes an unbalanced distribution of data, the model is more susceptible to situations where the recall for the minority class is zero or extremely low.

- To handle imbalanced data proposed a hybrid model SMOTE-NMA for majority and minority class.
- To properly balance the dataset for improving the model performance even more.

A significant portion of the predictions will be accurate for the majority class, while the features of the minority class will be ignored as noise in the data. The model will be highly biased as a result.

2. Literature Review

The sample size of an uncommon class (positive) is frequently a performance bottleneck in classification because of the class imbalance. However, "near-miss" positive examples, or negative but nearly positive instances, are occasionally common in real-world circumstances. Demonstrate that even in disaster predicting, where the real positive events are fairly few, the accuracy may be increased by getting finely honed label-like side-information "positivity" (for example, the river's water level) to identify near-miss situations from other negatives. Such side-information cannot be used in conventional cost-sensitive classification, and the limited size of the positive sample results in a significant estimation variance. In [2] Strategy adheres to the principles of learning using privileged information (LUPI), which uses side-knowledge for training without actually anticipating it. Theoretically demonstrate that, in return for increased bias and under the assumption that there are many near-miss positive examples, our strategy reduces the estimation variance. Extensive testing shows that the methodology typically performs better than or compares favorably to existing approaches.

Since conventional classification methods are built to handle balanced class distributions, learning from class-imbalanced data remains a frequent and difficult topic in supervised learning. For this objective, numerous algorithms have been developed, however the majority are complicated and frequently produce extra noise. In [3] work proposes a straightforward and efficient oversampling method based on k-means clustering and SMOTE (synthetic minority oversampling technique), which successfully addresses imbalances between and within classes while avoiding noise creation. The training data oversampled with the suggested strategy, according to comprehensive testing with 90 datasets, enhances classification outcomes. Additionally, k-means SMOTE routinely performs better than other well-liked oversampling techniques.

Discovering kNN using Euclidean without taking this correlation into account may result in discovering unrepresentative neighbours when a small number of attributes have higher correlation values than others. In [4] the paper introduces AWH-SMOTE (Attribute Weighted and kNN Hub on SMOTE), which advances SMOTE by improving noise and neighbour identification using attribute weighting and by improving the selective sampling approach by leveraging occurrence data in the kNN hub. A small number of occurrences in the kNN hub

leads to the generation of additional synthetic data, increasing the representation of minority data in risky regions. To assess AWH-SMOTE, nine open datasets from the Keel repository are used. Evaluation demonstrates that AWH-SMOTE outperforms alternative oversampling methods in terms of minority precision and minority f-measure for both pruned and unpruned conditions. When compared to other weighting methods for minority recall, minority precision, and minority f-measure, Information Gain as an attribute weighting method in AWH-SMOTE achieves the best performance in the unpruned situation.

The two-class unbalanced problem was the main focus of most classification techniques. Therefore, the multi-class imbalanced problem that exists in real-world domains must be solved. In [5] introduced a mechanism for classifying multi-class unbalanced data in the suggested work. There are two steps in this methodology: In the first phase, divided the original dataset into subsets of binary classes using binarization techniques (OVA and OVO). The SMOTE method is used in the second phase to balance the data by applying it to each subset of the imbalanced binary class. Finally, a Random Forest (RF) classifier is employed to achieve classification aim. MapReduce is specifically used to adapt the oversampling technique to big data so that it can handle any size data set that is required. To determine the effectiveness of the suggested strategy, an experimental study is conducted. Different datasets from the UCI repository were used for the experimental investigation, and the Apache Hadoop and Apache Spark platforms were used to implement the suggested solution. The outcomes demonstrate that the suggested strategy performs better than alternative methods.

In many real-world applications, very unbalanced data is present, which frequently makes machine-based processing challenging or impossible. The over- and under-sampling techniques help to address this problem, but they frequently have significant drawbacks. In [6] the study examines several class balancing techniques, focusing on results from under sampling. Additionally, a fresh approach is offered. The approach is focused on identifying and eliminating groups of examples from the majority class. As opposed to deleting individual cases or those from less-density areas, removing observations from high-density areas can result in less information loss. The distribution of examples is more evenly distributed with this method. The method's efficacy is proved through in-depth comparisons to various under sampling techniques using 18 open-source datasets. The outcomes of the experiments demonstrate that the suggested solution frequently enables greater performance than other evaluated techniques. The approach employed in this solution is based on the elimination of the nearest observations from the sample pool chosen as the k nearest neighbors of the majority class examples. However, the results of classification studies using the KNN_Order on eighteen datasets exceeded those from four under sampling approaches when they were compared in most cases.

3. Proposed Methodology

Proposed a methodology to handle the imbalance dataset for increasing the performance and do proper prediction of life threatening disease stage. Architecture of the model is given in fig. 3.1

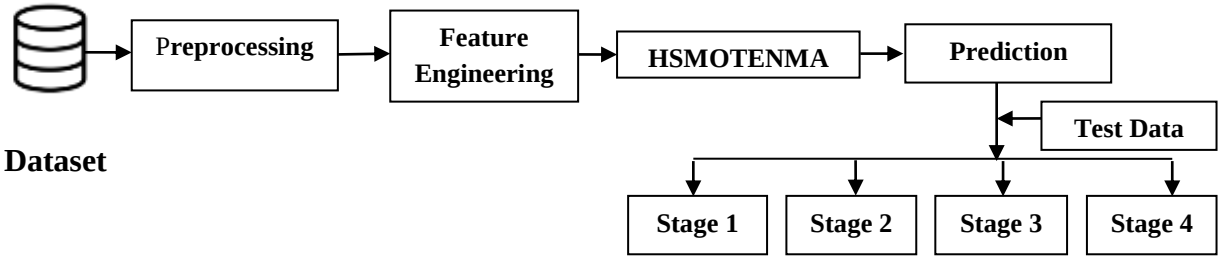

Figure: 3.1 Proposed Methodology

3.1.Dataset

Dataset used here is taken from UCI and kaggle [9] [8] website which contains the 615 entries and 19 columns as shown in fig. 3.2. Columns in the dataset are N_Days, Status, Drug, Age, Sex, Ascites, Hepatomegaly, Spiders, Edema, Bilirubin, Cholesterol, Albumin, Copper, Alk_Phos, SGOT, Tryglicerides, Platelets, Prothrombin, and Stage.

Status	Drug	Age	Sex	Ascites	Hepatomegaly	Spiders	Edema	Bilirubin	Cholesterol	Albumin	Copper	Alk_Phos	SGOT	Tryglicerides	Platelets	Prothrombin	Stage
D	D-penicillamine	21464	F	Y	Y	Y	Y	14.5	261.0	2.60	156.0	1718.0	137.95	172.0	190.0	12.2	4.0
C	D-penicillamine	20617	F	N	Y	Y	N	1.1	302.0	4.14	54.0	7394.8	113.52	88.0	221.0	10.6	3.0
D	D-penicillamine	25594	M	N	N	N	S	1.4	176.0	3.48	210.0	516.0	96.10	55.0	151.0	12.0	4.0
D	D-penicillamine	19994	F	N	Y	Y	S	1.8	244.0	2.54	64.0	6121.8	60.63	92.0	183.0	10.3	4.0
CL	Placebo	13918	F	N	Y	Y	N	3.4	279.0	3.53	143.0	671.0	113.15	72.0	136.0	10.9	3.0

Figure: 3.2 Dataset

3.2.Preprocessing

There are missing numbers for the majority of the numerical features, including bilirubin, prothrombin, triglycerides, etc. For handling the missing values the previous proposed MFSMOTE [1] imputation techniques applied.

3.3.Feature Engineering

Age values reconstructed to yearly observation diving by 365 days. The kurtosis values of the characteristics Bilirubin, Cholesterol, Copper, Alkaline Phosphatase, SGOT, Tryglicerides, and Prothrombin are more significant, and their distributions are severely skewed. They are therefore more likely to have outliers, as seen by their individual box plots as shown in fig. 3.3. Outliers are those data points which differs significantly from other observations present in given dataset. It might happen as a result of measurement variability and incorrect data point filling. To handle the outlier used IQR technique applied. Outliers are any observations that fall more than 1.5 IQR outside Q1 or raise more than 1.5 IQR outside Q3. Minitab's default method for locating outliers is this one.

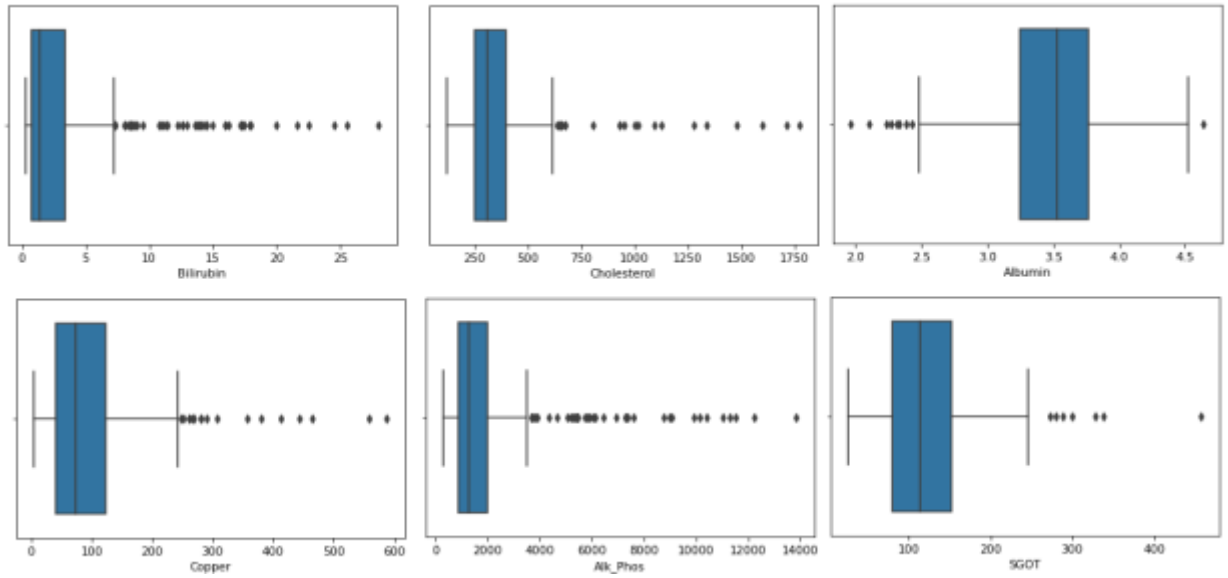

Figure: 3.3. Outlier

Nearly 38% of all patients had the third histologic stage of the Liver disease, which was followed closely by those with the fourth stage, which made up a proportion of roughly 35%. A little bit more than one-fifth of the total number of patients had the second stage of the cirrhosis disease. Patients with the first histologic stage of the disease, however, made up a very small portion of the total share as shown in fig. 3.4.

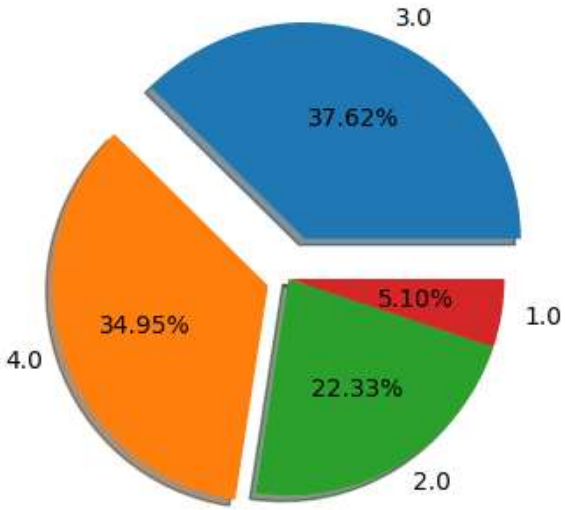

Figure: 3.4. Stage

Features like sex, edema, Ascites, Hepatomegaly, Spiders, Drug, and Stage are encoded with numerical values, at last the separated the independent predictor features and target labels. 'Status' and 'N_days' were removed since they were thought to be the source of data leakage.

3.3.1. HSMOTENMA

The target feature "Stage" has a very obvious class imbalance, with the third stage being the majority class and the first stage being the minority class as shown in fig. 3.5. The majority class falls under the "3.0" category while the minority class falls under the "1.0" label, creating a significant imbalance in the data. For classification, the data imbalance is a major issue. To

handle the imbalance combining oversampling and undersampling techniques to more evenly distribute the dataset can further improve model performance. Proposed HSMOTENMA an hybrid approach which combines the oversampling and undersampling algorithms.

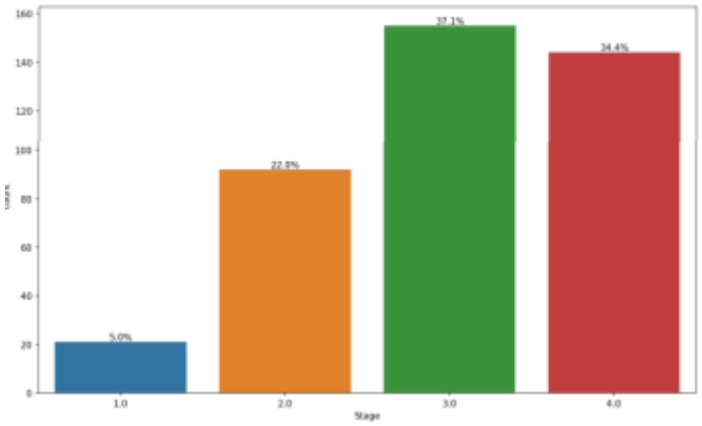

Figure: 3.5. Imbalanced Data

Oversampling done using the SMOTE (Synthetic Minority Oversampling Technique) and Undersampling done using NMA (Near Miss Algorithm). Each ones disadvantage overcome by one another. SMOTE (synthetic minority oversampling technique) is one of the most often utilised oversampling approaches to overcome the imbalance problem. By increasing minority class samples at random and duplicating them, it seeks to balance the distribution of classes. SMOTE creates new minority instances by combining minority instances that already exist. For the minority class, it creates virtual training records using linear interpolation. For each example in the minority class, one or more of the k-nearest neighbours are randomly chosen to serve as these synthetic training records. Following the oversampling procedure, the data is rebuilt and can be subjected to several categorization models. A method of undersampling is called NearMiss. By randomly removing examples from the dominant class, it seeks to balance the distribution of classes. We eliminate instances of the majority class when instances of two separate classes are relatively close to one another in order to enhance the distances between the two classes [10]. This facilitates the process of categorization. Near-neighbor strategies are frequently employed to prevent the problem of information loss in the majority of under-sampling techniques.

Algorithm: HSMOTENMA

- Initial: (Begin with SMOTE) Select arbitrary information from the minority class.
- Step 2: Determine the Manhattan distance between the k closest neighbours of the random data and it.
- Step 3: The uneven percentage is used to determine the sample rate N. N instances are randomly chosen from each of each's k-nearest neighbours and used to build the set for each (i.e. x1, x2,... xn).
- Step 4: Continue performing steps 1-2 until the desired percentage of the minority class is reached. (SMOTE ends here)
- Step 5: (Begin NMA) The procedure starts by measuring the separations between every member of the majority class and every member of the minority class. In this instance, the majority class is to be undersampled.

- Step 6: Next, the n instances of the majority class that are closest to the minority class members are chosen.
    - The observation and its K-nearest neighbour are removed from the dataset if the class of the observation and the majority class from the observation's K-nearest neighbour are different.
    - The closest method will produce k*n instances of the majority class if there are k instances of the minority class.
- Step 7: Repeat steps 2 and 3 as necessary to achieve the desired percentage of each class. (ENN ends)

3.4.Prediction

Dataset trained on the base models Logistic Regression (LR), Ridge Classifier (RC), and Gausssian NB (GNB). Base models further tuned Hyperparameter Tuning and "RepeatedStratifiedKFold" cross validation technique.

4. Result and Discussion

For evaluating the performance of the models precision, recall, f1-score, fitting time, and test accuracy is used. The base model with imbalanced data and without imbalanced data analyzed to analyze the performance of the proposed model.

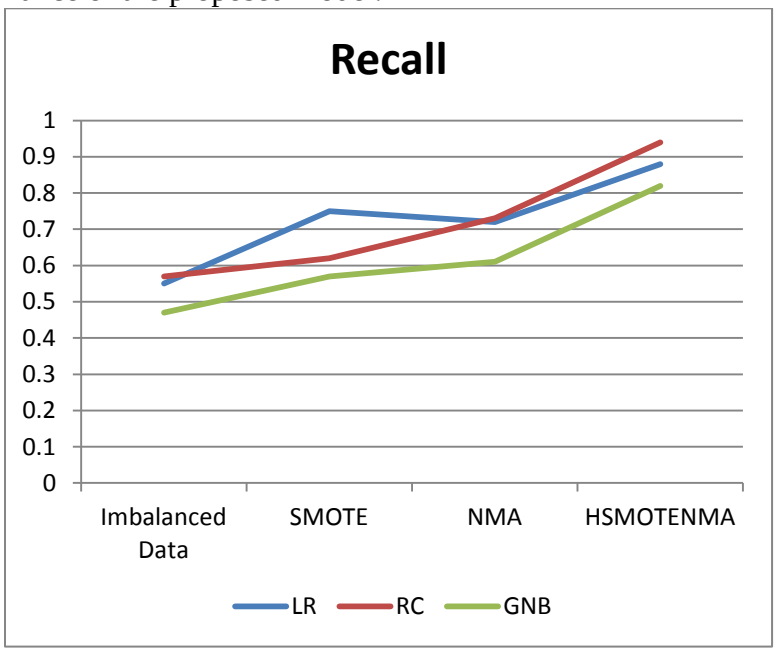

Figure: 4.1 Recall

From fig.4.1 it is very clear the proposed model HSMOTENMA having the highest recall of 0.88, 0.94, and 0.82 with the base models LR, RC, and GNB respectively. It is also understand that imbalanced data gives a very low recall rate of 0.55, 0.57, and 0.47 with base models LR, RC, and NGB respectively.

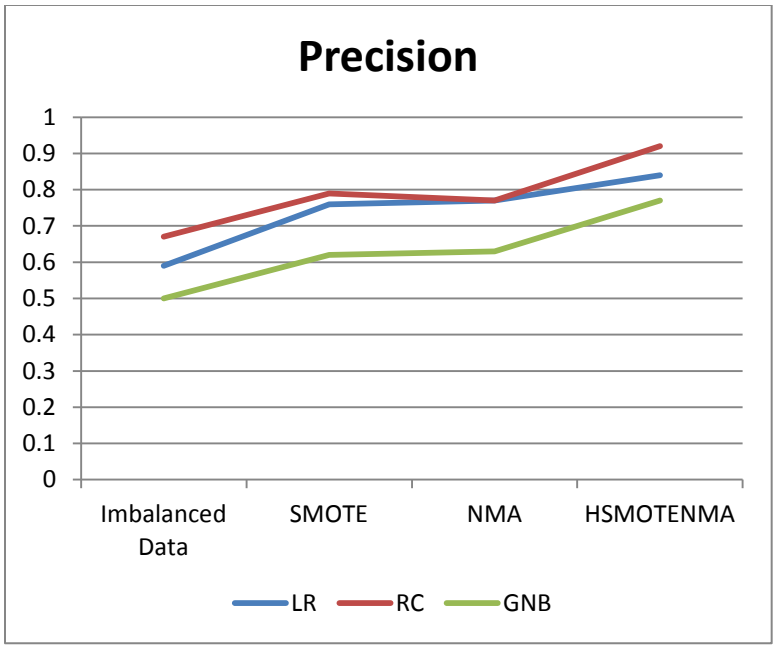

Figure: 4.2 Precision

From fig.4.2 it is very clear the proposed model HSMOTENMA having the highest precision of 0.84, 0.92, and 0.77 with the base models LR, RC, and GNB respectively. It is also understand that imbalanced data gives a very low precision rate of 0.59, 0.67, and 0.50 with the base models LR, RC, and NGB respectively.

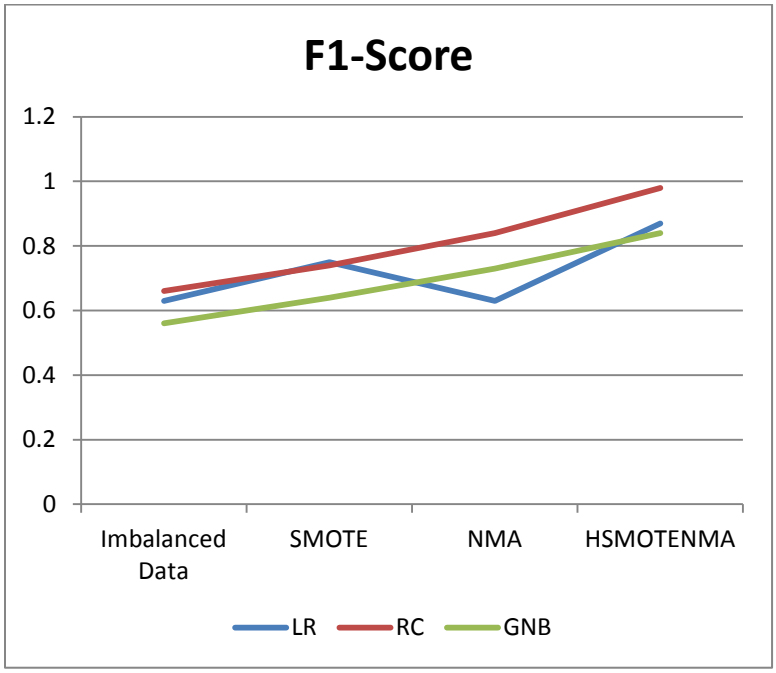

Figure: 4.3 F1-Score

From fig.4.3 it is very clear the proposed model HSMOTENMA having the highest f1-score of 0.87, 0.98, and 0.84 with the base models LR, RC, and GNB respectively. It is also understand

that imbalanced data gives a very low f1-score of 0.63, 0.66, and 0.56 with the base models LR, RC, and NGB respectively.

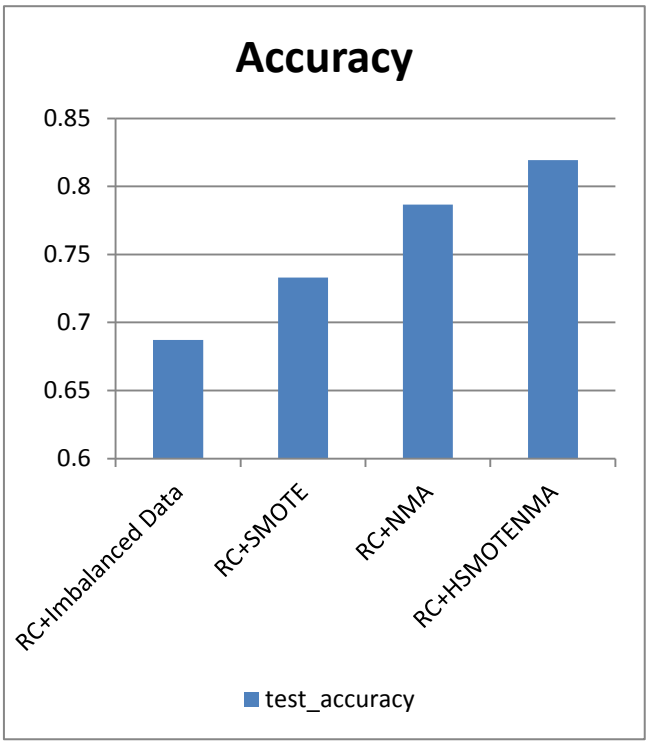

Figure: 4.4 Accuracy

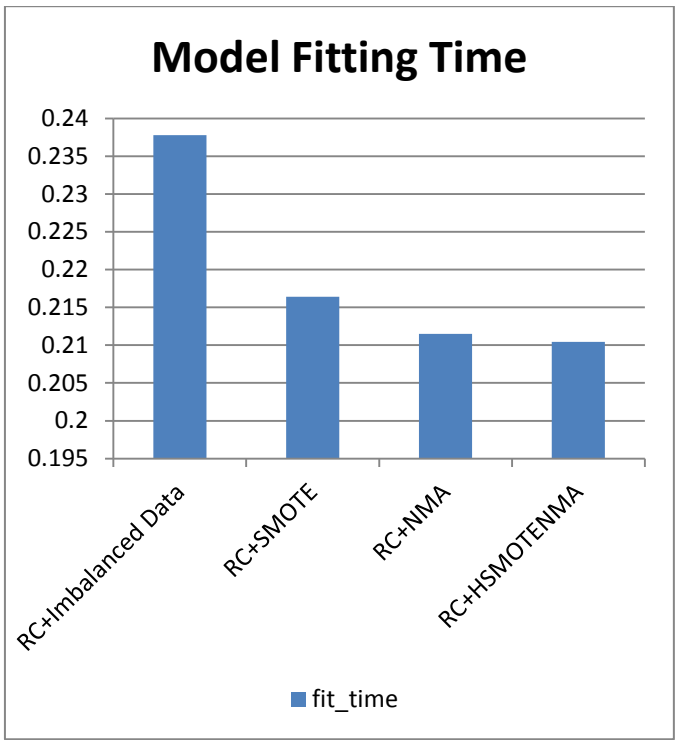

Figure: 4.5 Model Fitting Time

Base model RC performs better with the proposed model HSMOTENMA. So the base model RC is choice for further prediction on the Test data. On test data it yields a highest accuracy of 81% as shown in fig. 4.4 and it takes less time to fitting data into model of 0.2103 seconds as shown in fig. 4.5.

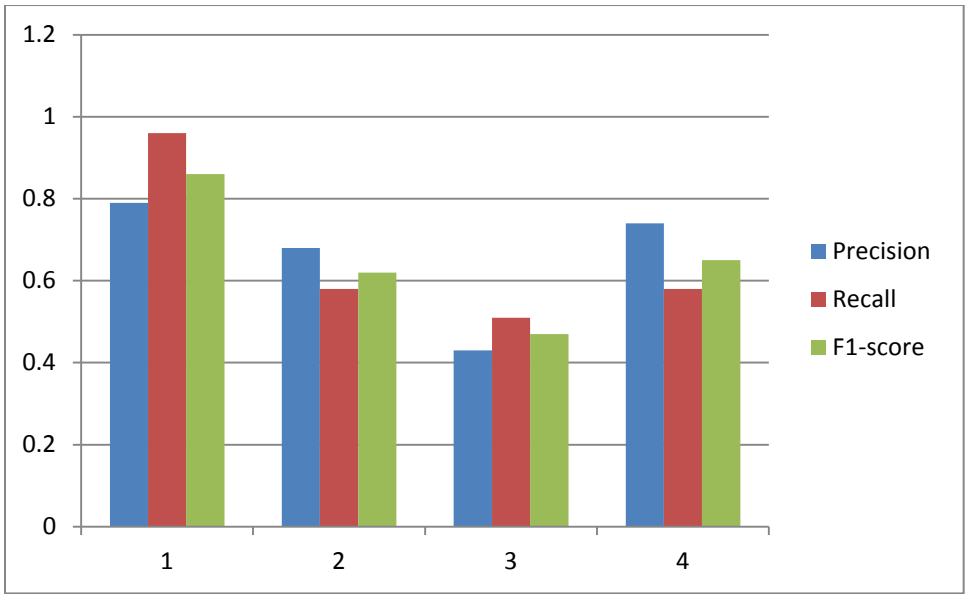

Figure: 4.6. Prediction Using Ridge Classifier with HSMOTENMA

Selected model RC+HSMOTENMA performed the prediction on the test data. Prediction of different stage of Liver disease determined (as shown in fig. 4.6). The prediction evict that it handles a balanced data on both Majority and Minority Classes.

5. Conclusion

Liver Disease prediction to identify the stage of disorder. To find the disease in early stage will save a life. While preparing the model with the available dataset we found a classification issue of determining stages. There is misclassification of stages. Even the stages of the dataset are imbalanced. To overcome the problem of handling the imbalanced data proposed a hybrid approach HSMOTENMA combining oversampling and under sampling technique. The proposed model yields high accuracy and recall compare with the handling individually. Nearly the recall rate reaches 0.92 and accuracy of 81%. Analyzed the performance of the proposed model using base classifiers like LR, RC, and GNB. RC performs well with the proposed model. Further RC+HSMOTENMA used for predicting the test data.

Reference:

1. Sindhiya, "Iterative Imputation Preprocessing Techniques for handling Missing Data in Liver Disease Prediction", ,Vol. , Is. , 2023.
2. Tanimoto, A., Yamada, S., Takenouchi, T., Sugiyama, M., & Kashima, H. (2022, September). Improving imbalanced classification using near-miss instances. *Expert Systems with Applications*, *201*, 117130. https://doi.org/10.1016/j.eswa.2022.117130
3. Douzas, G., Bacao, F., & Last, F. (2018, October). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Information Sciences*, *465*, 1–20. https://doi.org/10.1016/j.ins.2018.06.056
4. Fahrudin, Buliali, & Fatichah. (2019). ENHANCING THE PERFORMANCE OF SMOTE ALGORITHM BY USING ATTRIBUTE WEIGHTING SCHEME AND NEW SELECTIVE SAMPLING METHOD FOR IMBALANCED DATA SET. *International Journal of Innovative Computing, Information and Control*, *15*(2), 223–244. https://doi.org/10.24507/ijicic.15.02.423
5. Bhagat, R. C., & Patil, S. S. (2015, June). Enhanced SMOTE algorithm for classification of imbalanced big-data using Random Forest. *2015 IEEE International Advance Computing Conference (IACC)*. https://doi.org/10.1109/iadcc.2015.7154739
6. Bach, M., Werner, A., & Palt, M. (2019). The Proposal of Undersampling Method for Learning from Imbalanced Datasets. *Procedia Computer Science*, *159*, 125–134. https://doi.org/10.1016/j.procs.2019.09.167
7. Mondal, D., Das, K., & Chowdhury, A. (2022, January 28). Epidemiology of Liver Diseases in India. Clinical Liver Disease, 19(3), 114–117. https://doi.org/10.1002/cld.1177
8. fedesoriano. (August 2021). Cirrhosis Prediction Dataset. Retrieved [Date Retrieved] from https://www.kaggle.com/fedesoriano/cirrhosis-prediction-dataset.
9. UCI data set, http://archive.ics.uci.edu/ml/machine-learning-databases/00225/
10. Feng, S., Keung, J., Zhang, P., Xiao, Y., & Zhang, M. (2022, February). The impact of the distance metric and measure on SMOTE-based techniques in software defect prediction. Information and Software Technology, 142, 106742. https://doi.org/10.1016/j.infsof.2021.106742
11. Srivastava, A., Kumar, V. V., R, M. T., & Vivek, V. (2022, April 21). Automated Prediction of Liver Disease using Machine Learning (ML) Algorithms. 2022 Second

International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT). https://doi.org/10.1109/icaect54875.2022.9808059