2581-4575
International Journal For Recent Developments in Science & Technology
Crossref
A Peer Reviewed Research Journal

PREDICTION OF LIVER DISEASE WITH RANDOM FOREST CLASSIFIER THROUGH SMOTE-ENN BALANCING

M. Mahipal Reddy[1], A Deekshitha[2], Mamidi Yashwanth Babu[2], Retineni Naveen[2], Ragula Vaj Bharath[2], Macharla Srujan[2]

[1]Assistant Professor, [2]UG Student, [1,2]Department of Computer Science Engineering
[1,2]Malla Reddy Engineering College and Management Science, Kistapur, Medchal-50140l, Hyderabad, Telangana, India

Abstract

The human liver is the largest internal organ of the body and liver disease is among the critical diseases that affect the normal, healthy stature of a human due to various reasons. There are various types of liver disease, namely fatty liver, cirrhosis, hepatitis, chronic liver disease, liver cancer and liver tumor, etc. Excess triglyceride fat accumulation leads to fatty liver. This work develops machine learning algorithms to improve Liver Disease prediction by experimenting with various data balancing techniques such as Synthetic Minority Oversampling Technique (SMOTE), SMOTE+ Encoded Nominal and Continuous (ENC), SMOTE+ Edited Nearest Neighbourhood (ENN), SMOTE+TOMEK, KMEANS SMOTE and SVM SMOTE and in algorithms Random Forest giving best accuracy with SEMOT-ENN balancing technique. Further, correlation and skewness techniques are used to clean the dataset and to reduce features. Then, PCA (principal component analysis) algorithm is used and to normalize features. Finally, prediction operation is carried out by Random Forest classifier with SEMOTE-ENN balancing technique. Liver Disease dataset which can be download from UCI or KAGGLE website has used. All techniques performance is evaluated in terms of accuracy, precision and confusion matrix.

Keywords: Liver disease, Synthetic Minority Oversampling Technique, Encoded Nominal and Continuous, Edited Nearest Neighbourhood, SMOTE-TOMEK, KMEANS-SMOTE, SVM-SMOTE

1. INTRODUCTION

The liver is an important organ of the human body, and it is located beneath the rib cage in the right upper abdomen. It removes toxins from the body and maintains healthy blood sugar level in the body. Though body organs have self-healing capacity, over consumption of alcohol and exposure to impure air and water affects the liver which leads to higher rate of Liver failure. Liver transplantation is the solution but with higher cost and lower rate of success. Identifying the liver damage at the earliest can reduce the chance of liver failure. The machine-learning model is capable of predicting diseases, based on a data set, which is built in combination of key health parameters of a person with diseases and without diseases. For building models, an effective data set is needed, with proper representation of disease classifications. Problems with the liver are difficult to detect early on since it will continue to operate normally even if it is partially destroyed. The chances of a patient surviving a liver disease are better if they are diagnosed early. Indians are at a higher risk of liver failure. India is anticipated to become the World Capital for Liver Diseases by 2025. In India, a deskbound lifestyle, increased alcohol intake, and smoking are all factors contributing to the prevalence of liver infection. There are over a hundred different forms of liver infections. As a result, inventing a machine that would aid in disease identification will be extremely beneficial in the medical industry. These technologies will assist

physicians in making accurate patient decisions, and with the use of automatic classification tools for liver illnesses (likely mobile enabled or web enabled), the patient wait at liver experts such as endocrinologists will be reduced.

To detect disease, healthcare professionals need to collect samples from patients which can cost both time and money. Often, more than one kind of test or many samples are needed from the patient to accumulate all the necessary information for a better diagnosis. The most routine tests are urinalysis, complete blood count (CBC), and comprehensive metabolic panel (CMP). These tests are generally less expensive and can still be very informative. The liver has many functions such as glucose synthesis and storage, detoxification, production of digestive enzymes, erythrocyte regulation, protein synthesis, and various other features of metabolism. Chronic liver diseases include chronic hepatitis, fibrosis, and cirrhosis. Hepatitis can occur from viral infection (e.g., hepatitis c virus) or auto-immune origin. Inflammation from hepatitis infection can cause tissue damage and scarring to occur in the liver. Moderate scarring is classified as fibrosis, while severe liver damage/scarring is classified as cirrhosis. Fibrosis and cirrhosis can also occur from alcoholism and non-alcoholic fatty liver disease. When liver disease is diagnosed at an earlier stage, in between infection and fibrosis but before cirrhosis, liver failure can be avoided. Tests, such as a CMP and biopsy, can be conducted to diagnose all forms of liver disease. A CMP with a liver function panel can detect albumin (ALB), alkaline phosphatase (ALP), alanine amino-transferase (ALT), aspartate amino-transferase (AST), gamma glutamyl-transferase (GGT), creatine (CREA), total protein (PROT), and bilirubin (BIL). Diagnosis of a certain liver disease and discovery of its origin are made by interpreting the patterns and ratios of circulating liver-associated molecules measured with the CMP test and compared to values normalized with a patient's age, sex, and BMI. Aminotransferases, AST, and ALT are enzymes that participate in gluconeogenesis by catalysing the reaction of transferring alpha-amino groups to ketoglutaric acid groups. AST is found in many tissue types and is not as specific to the liver but may denote secondary non-hepatic causes of liver malfunction.

Classification techniques are very popular in various automatic medical diagnoses tools. Problems with liver patients are not easily discovered in an early stage as it will be functioning normally even when it is partially damaged [1]. An early diagnosis of liver problems will increase patients survival rate. Liver disease can be diagnosed by analyzing the levels of enzymes in the blood [2]. Moreover, now a day's mobile devices are extensively used for monitoring humans' body conditions. Here also, automatic classification algorithms are needed. With the help of Automatic classification tools for liver diseases (probably mobile enabled or web enabled), one can reduce the patient queue at the liver experts such as endocrinologists. Michael J Sorich [3] reported that SVM classifier produces best predictive performance for the chemical datasets. Lung-Cheng Huang reported that Naïve Bayesian classifier produces high performance than SVM and C 4.5 for the CDC Chronic fatigue syndrome dataset [5]. Paul R Harper [4] reported that there is not necessary a single best classification tool but instead the best performing algorithm will depend on the features of the dataset to be analyzed.

2. LITERATURE SURVEY

Shackel et. al [6] used methods of transcriptome analysis, especially gene array analysis, focusing on publications utilizing these methods to understand human liver disease. Additionally, they have outlined the relationship between transcript and protein expressions as well as summarizing what is known about the variability of the transcriptome in non-diseased liver tissue. The approaches covered include gene array analysis, serial analysis of gene expression, subtractive hybridization and differential display. The discussion focuses on primate whole organ studies and in-vitro cell culture systems utilized. It is now

clear that there are a vast number research opportunities for transcriptome analysis of human liver disease as we attempt to better understand both non-diseased and disease hepatic mRNA expression. They conclude that hepatic transcriptome analysis has already made significant contributions to the understanding of human liver pathobiology.

Lin et. al [7] Liver disease, the most common disease in Taiwan, is not easily discovered in its initial stage; early diagnosis of this leading cause of mortality is therefore highly important. The design of an effective diagnosis model is therefore an important issue in liver disease treatment. This work accordingly employs classification and regression tree (CART) and case-based reasoning (CBR) techniques to structure an intelligent diagnosis model aiming to provide a comprehensive analytic framework to raise the accuracy of liver disease diagnosis.

VenkataRamana et. al [8] popular Classification Algorithms were considered for evaluating their classification performance in terms of Accuracy, Precision, Sensitivity and Specificity in classifying liver patient's dataset. Accuracy, Precision, Sensitivity and Specificity are better for the AP Liver Dataset compared to UCLA liver datasets with all the selected algorithms. This can be attributed to a greater number of useful attributes like Total bilirubin, Direct bilirubin, Indirect bilirubin, Albumin, Gender, Age and Total proteins are available in the AP liver dataset compared to the UCLA dataset. The common attributes for AP liver data and Taiwan data are Age, Sex, SGOT, SGPT, ALP, Total Bilirubin, Direct Bilirubin, Total Proteins and Albumin are crucial in deciding liver status. . With the selected dataset, KNN, Back propagation and SVM are giving better results with all the feature set combinations.

Kumar and Sahoo et. al [9] proposed a rule-based classification model with machine learning techniques for the prediction of different types of Liver diseases. A dataset was developed with twelve attributes that include the records of 583 patients in which 441 patients were male and rests were female. Support Vector Machine (SVM), Rule Induction (RI), Decision Tree (DT), Naive Bayes (NB) and Artificial Neural Network (ANN) data mining techniques with K-cross fold technique are used with the proposed model for the prediction of liver diseases. The performance of these data mining techniques are evaluated with accuracy, sensitivity, specificity and kappa parameters as well as statistical techniques (ANOVA and Chi square test) are used to analyze the liver disease dataset and independence of attributes.

Sontakke et. al [10] explores 2 methodologies in chronic liver disease prediction. Liver disease is especially difficult to diagnose given the subtle nature of its symptoms. Of the 2,626,418 deaths reported in the United States for 2014, chronic liver disease accounted for nearly 38,170 deaths. Prediction by means of computers will continue to grow in importance. This work explored 2 possibilities of machine learning models that can improve predictive power. The molecular biology approach is often affected by diet, age, and ethnicity. The chemical approach is a surer method of prediction. However in all eventuality, research in the direction of molecular biology can help unravel the secrets to human anatomy which will help save lives.

Gogi and V. G. N. et. al [11] makes use of the lab test reports of the patients who has undergone Liver Function Test. MATLAB2016 is used and developed a model by applying classification algorithms SVM, Logistic Regression and Decision tree. Logistic Regression gave high accuracy of 95.8%.Various predictors are tested by plotting graph that determined the existence of disorder in liver.

Javad Hassannataj et. al [12] to select significant features by comparing data mining models to predict liver disease based on an extraction, loading, transformation, analysis (ELTA) approach for correct

diagnosis. Hence, the data mining models are compared based on the ELTA approach, such as random forest, Multi-Layer Perceptron (MLP) neural network, Bayesian networks, Support Vector Machine (SVM), and Particle Swarm Optimization (PSO)-SVM. Among these models, the PSO-SVM model has the best performance regarding the criteria of specificity, sensitivity, accuracy, Area under the Curve (AUC), F-measure, precision, and False Positive Rate (FPR). Furthermore, a 10-fold cross-validation method for evaluation of models is used so that the models were evaluated on a liver disease dataset. The average of estimated accuracy was calculated as 87.35%, 78.91%, 66.78%, 76.51% and 95.17% for Random forest, MLP Neural network, Bayesian network, SVM and PSO-SVM models, respectively. Regarding the mentioned evaluation criteria, we obtained the highest performance of accuracy with the least number of features through the hybrid PSO-SVM-based optimized model.

Ambesange et. al [13] build the machine-learning model, Indian Liver Patient Dataset (ILPD) hosted at UCI.edu is used, which is based on Indian patient and Random Forest (RF) algorithm is used to predict the disease with different preprocessing techniques. Data set is checked for skewness, outliers and imbalance using univariate and bivariate analysis and then suitable algorithms used to remove outliers and various oversampling and under sampling techniques are used to balance the data. Further refinement of model is done through hyper parameter tuning using grid search and feature selection. The final model provides 100% accuracy and also good score across different metrics.

Kuzhipallil et. al [14] compares various classification models and visualization techniques used to predict liver disease with feature selection. Outlier detection is used to find out the extreme deviating values and they are eliminated using isolation forest. The performance is measured in terms of accuracy, precision, recall f-measure and time complexity. The results of various classifiers are obtained by using proposed feature selection algorithm. From the experiments and comparative analysis, it increases classification accuracy and also leads to reduction in classification time and hence aids in the prediction of the disease more efficiently.

Deshmukh et. al [15] presented a deep-learning-based framework for the segmentation of vacuoles in liver images of Wistar rat and study the correlation of automated quantification with expert pathologist's manual evaluation. To address the issue of misclassification of lumina (vascular and bile duct) as large vacuoles, we propose a selective tiling technique to generate tiles that include complete lumina and large vacuoles. A binary encoder-decoder convolution neural network is trained to detect individual vacuoles. They report a sensitivity of 85% and specificity of 98%. Furthermore, the diameter and roundness of the segmented vacuoles are estimated with an error of less than 8%, which supports the high potential of our method in drug development process.

3. PROPOSED SYSTEM

The human liver is the largest internal organ of the body and liver disease is among the critical diseases that affect the normal, healthy stature of a human due to various reasons. There are various types of liver disease, namely fatty liver, cirrhosis, hepatitis, chronic liver disease, liver cancer and liver tumor, etc. Excess triglyceride fat accumulation leads to fatty liver. Hepatitis virus infection in the liver is developed due to excessive consumption of alcohol, detrimental food habits, etc.

Hepatitis can result in acute and chronic infection. Cirrhosis is fibrous tissue that replaces the dead liver cells with fibroid. Metastatic tumors are cancerous tumors in the liver that spread from cancer affected in other organs. Chronic liver disease (CLD) is estimated worldwide with figures up to 844 million people and has a mortality rate of about two million per year. The World Health Organization (WHO) states that death tolls worldwide rose to 50 million per year over two decades due to cirrhosis and liver

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

cancer. In 2015, deaths caused by liver diseases due to alcohol were reported to be an age-standardized rate of 14.2 deaths per 100,000 population by the global studies made in the United Kingdom. Figure 1 shows the block diagram of proposed methodology. Here, liver disease dataset considers a format of CSV file then preprocessing operation is performed to remove the missing spaces and fill the empty characters for that this preprocessing operation also normalizes the entire dataset with uniform sizing's next feature from dataset are extracted by using PCA and Skewness properties. Here, PCA holds the important feature properties with reducing dimensions then dataset balancing between two classes is archives by using SMOTE-ENN balancing technique these SMOTE-ENN technique balances the two classes of dataset. Finally, random forest classification mechanism is applied to perform the different types of prediction operations.

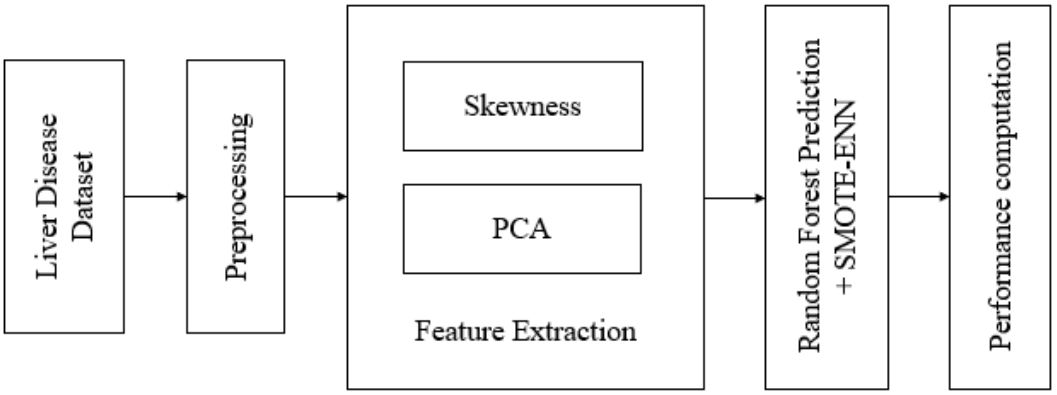


Figure 1. Block Diagram of Proposed system

3.1 Dataset

This data set contains 416 liver patient records and 167 non liver patient records collected from North East of Andhra Pradesh, India. The "Dataset" column is a class label used to divide groups into liver patient (liver disease) or not (no disease). This data set contains 441 male patient records and 142 female patient records. Any patient whose age exceeded 89 is listed as being of age "90".

Columns**:**

- Age of the patient
- Gender of the patient
- Total Bilirubin
- Direct Bilirubin
- Alkaline Phosphotase
- Alamine Aminotransferase
- Aspartate Aminotransferase
- Total Protiens
- Albumin
- Albumin and Globulin Ratio

Dataset: field used to split the data into two sets (patient with liver disease, or no disease)

3.2 Preprocessing

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data preprocessing task.

Need of Data Preprocessing: A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

Feature Scaling: Feature scaling is the final step of data preprocessing in machine learning. It is a technique to standardize the independent variables of the dataset in a specific range. In feature scaling, we put our variables in the same range and in the same scale so that no variable dominates the other variable. A machine learning model is based on Euclidean distance, and if we do not scale the variable, then it will cause some issue in our machine learning model. Euclidean distance is given as:

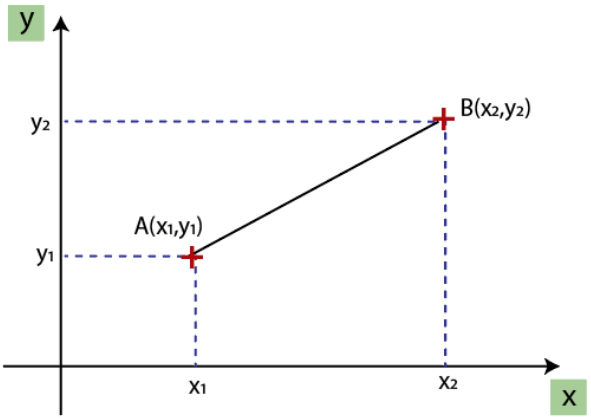


Figure 2. Feature scaling

If we compute any two values from age and salary, then salary values will dominate the age values, and it will produce an incorrect result. So to remove this issue, we need to perform feature scaling for machine learning.

3.3 Splitting the Dataset

In machine learning data preprocessing, we divide our dataset into a training set and test set. This is one of the crucial steps of data preprocessing as by doing this, we can enhance the performance of our machine learning model. Suppose if we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models. If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as:

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

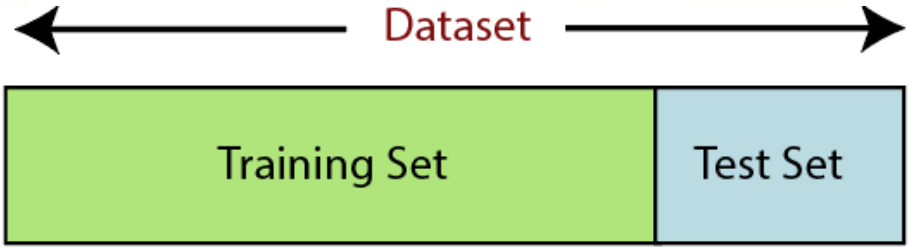


Figure 3. Splitting the dataset

Training Set: A subset of dataset to train the machine learning model, and we already know the output.

Test set: A subset of dataset to test the machine learning model, and by using the test set, model predicts the output.

3.4 PCA feature reduction

The Principal Component Analysis is a popular unsupervised learning technique for reducing the dimensionality of data. It increases interpretability yet, at the same time, it minimizes information loss. It helps to find the most significant features in a dataset and makes the data easy for plotting in 2D and 3D. PCA helps in finding a sequence of linear combinations of variables.

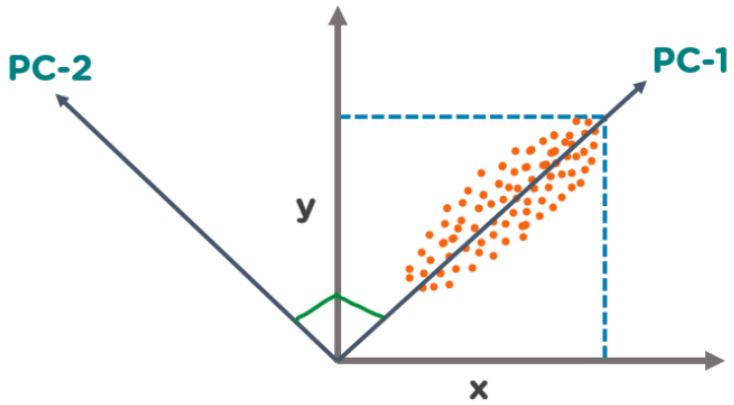


Figure 4. PCA analysis

In the above figure, we have several points plotted on a 2-D plane. There are two principal components. PC1 is the primary principal component that explains the maximum variance in the data. PC2 is another principal component that is orthogonal to PC1.

Figure 5. Applications of PCA in Machine Learning

- PCA is used to visualize multidimensional data.

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

- It is used to reduce the number of dimensions in healthcare data.
- PCA can help resize an image.
- It can be used in finance to analyze stock data and forecast returns.
- PCA helps to find patterns in the high-dimensional datasets.

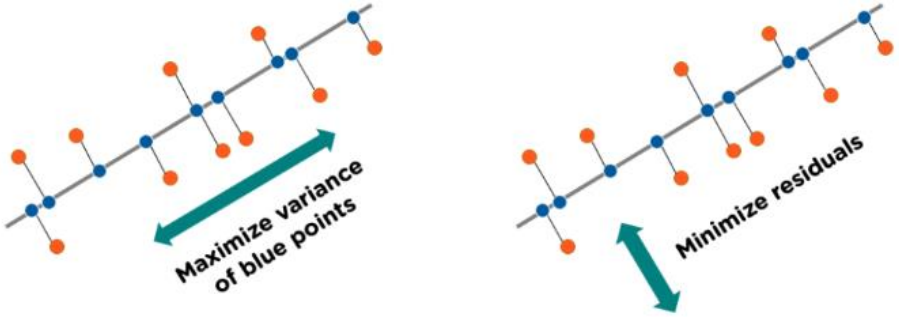


Figure 6. PCA working

Step 1: Normalize the data: Standardize the data before performing PCA. This will ensure that each feature has a mean = 0 and variance = 1.

$$Z = \frac{x - \mu}{\sigma}$$

Step 2: Build the covariance matrix: Construct a square matrix to express the correlation between two or more features in a multidimensional dataset.

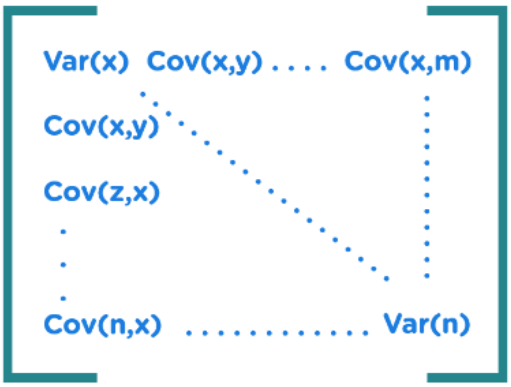


Figure 7. Covariance matrix formulation

Step 3: Find the Eigenvectors and Eigenvalues: Calculate the eigenvectors/unit vectors and eigenvalues. Eigenvalues are scalars by which we multiply the eigenvector of the covariance matrix.

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

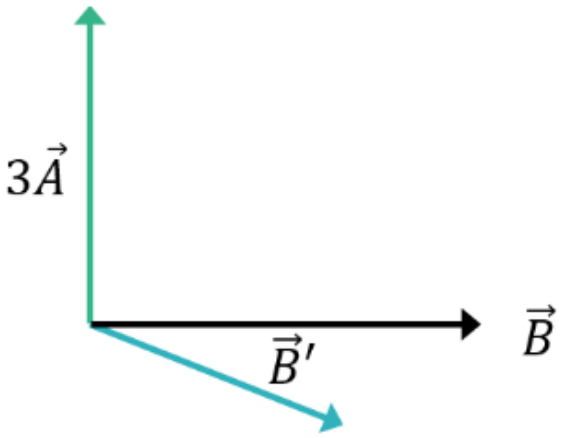


Figure 8. PCA dimension reduction

Step 4: Sort the eigenvectors in highest to lowest order and select the number of principal components.

3.4 SMOTE-ENN

Since there is an imbalance in sample size between non-LD and LD samples, we apply the synthetic minority oversampling technique (SMOTE) and the edited nearest neighbor (ENN) technique to artificially increase the LD samples. The main idea of SMOTE is to create numerous new cases of minority class by randomly choosing a near neighbor of the minority class and interpolating as described. First, for each sample of minority class Xp, its k nearest neighbors from other samples of minority class are taken. Subsequently, the minority class sample Xq among the k neighbors, is randomly selected. In the last step, XNew is generated as the synthetic sample by interpolating Xp and Xq:

$$XNew = Xp + rand\,(0,1)$$

where rand (0, 1) represents a generated random number between 0 and 1. From a geometric viewpoint, the process of utilizing SMOTE can be considered an interpolation between two LD samples. The decision space for the LD samples is thus magnified. The SMOTE method can balance the number of each category. However, it may cause the generated minority class samples and the original majority class samples to overlap so that they cannot be discriminated well.

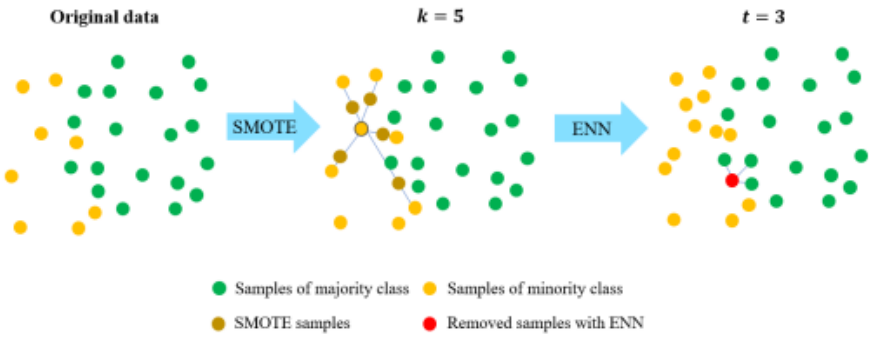


Figure 9. Illustration of SMOTE+ENN.

To solve this problem, we use the SMOTE+ENN method. In the training set, the ENN method is used to find t neighboring cases for the minority lass cases (XNew) generated by SMOTE. If these t cases

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

belong to the cases of the majority class, the cases of the minority class will be deleted. In this case, the SMOTE+ENN method can make the boundary of each class clearer and allows the proposed machine learning classifiers to have higher prediction performances on the unknown LD samples. In this study, the SMOTE+ENN method as illustrated in Figure 9, is utilized in the process of 2-fold cross-validation

Adaptive Synthetic Sampling: Adaptive Synthetic Sampling adaptively generates different numbers of sampling samples according to data distribution. The basic flow of the algorithm is below:

Step 1: Calculate the number of samples to be synthesized, as follows: G = (ml − ms) × β, where ml is the number of majority samples, andms is the number of minority samples. If β = 1, the number of positive and negative samples is the same after sampling, indicating that the data is balanced at this time.

Step 2: Calculate the number of K nearest neighbor value of each minority sample, 1 is the number of majority samples in the K neighbors, the formula is as follows: ri = 1i/K, where 1i

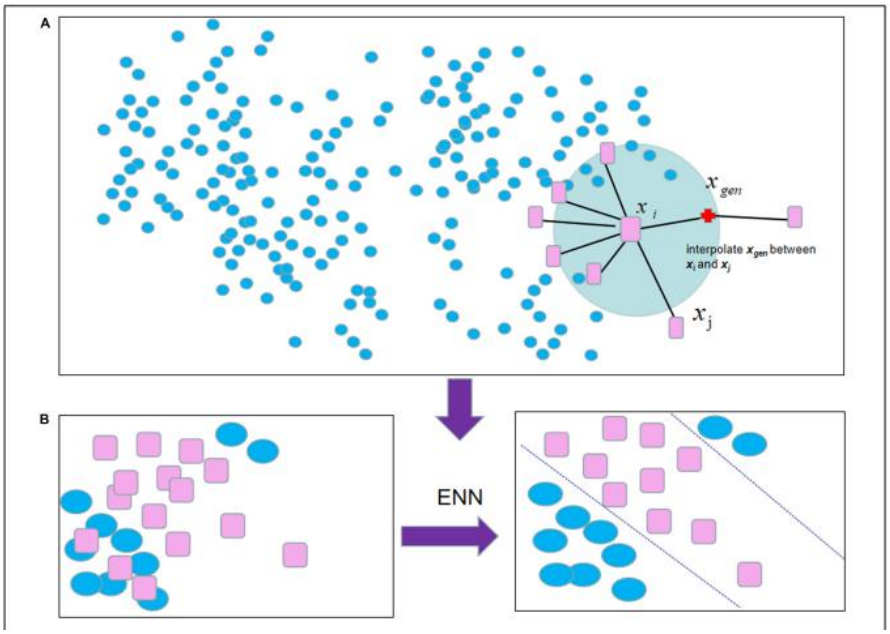
is the number of majority samples in K nearest neighbors, i = 1,2,3....., ms

Step 3: To normalize ri , the formula is rˆ = ri/ Pms i=1 ri

Step 4: According to the sample weights, calculate the number of new samples that need to be generated for every minority sample. The formula is g = ˆr × G. Select one sample from the K neighbors around each data with the label "1" to be synthesized, calculate the number to be generated according to g the formulas i = xi + (xzi − xi) × λ

where si is the synthetic sample, xi is the ith minority samples, and xzi is a random number of the minority sample λ ∈ [0,1] selected from the K nearest neighbors of xi

Apart from using a single under-sampling or over-sampling method, two resampling methods can be combined. For example, SMOTE-ENN, ENN is an under-sampling method focusing on eliminating noise samples, which is added to the pipeline after SMOTE to obtain cleaner combined samples. For each combined sample, its nearest-neighbors are computed according to the Euclidean distance. These samples will be removed whose most KNN samples are different from other classes



International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

Figure 10: The process of SMOTE-ENN algorithm: (a) SMOTE selected each sample from the minority samples successively as the root sample for the synthesis of the new sample. (b) The following result was obtained by employing ENN to eliminate noise samples when the process of SMOTE is caused.

3.5 Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML.

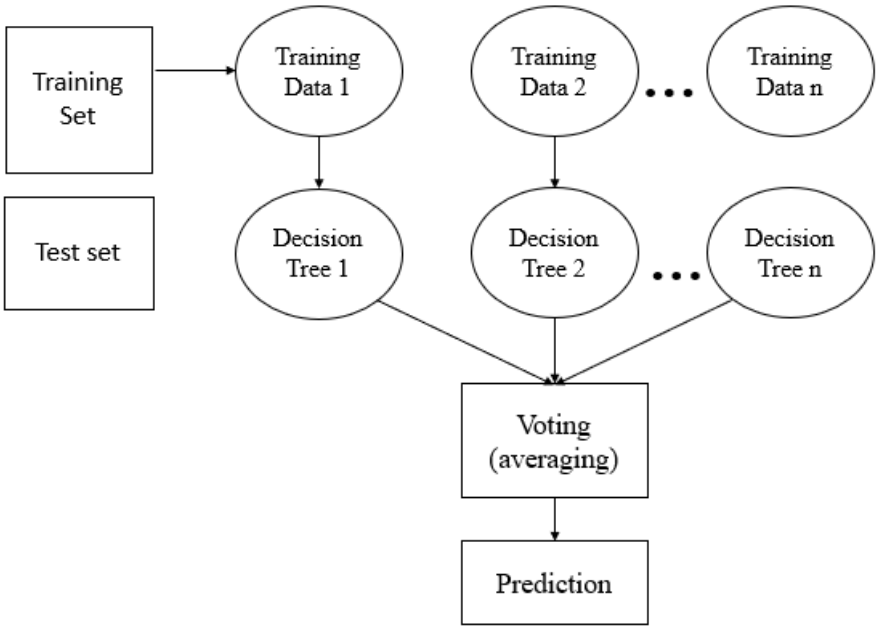


Figure 11. Random Forest algorithm

It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. Random Forest algorithm

Step 1: In Random Forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

Important Features of Random Forest

- **Diversity**- Not all attributes/variables/features are considered while making an individual tree, each tree is different.

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

- **Immune to the curse of dimensionality**- Since each tree does not consider all the features, the feature space is reduced.
- **Parallelization**-Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.
- **Train-Test split**- In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.
- **Stability**- Stability arises because the result is based on majority voting/ averaging.

4. RESULTS AND DISCUSSION

This section gives the detailed analysis of simulation results implemented using "python environment". Further, the performance of proposed method is compared with existing methods using same dataset.

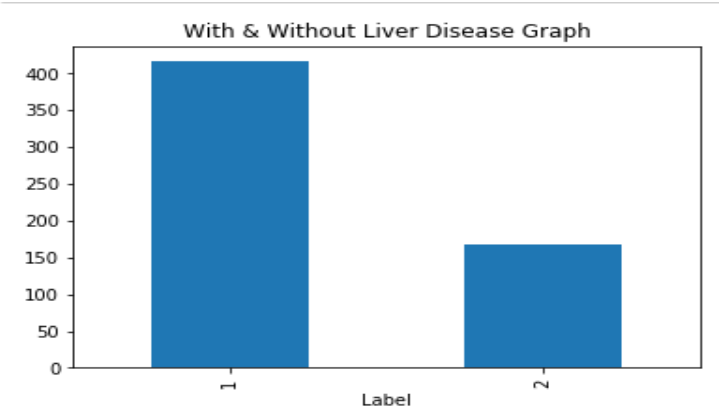


Figure 12. No of Diseases.

In Figure 12, x-axis represents 1 (no disease) and 2 (disease present) and y-axis contains number of records and in above screen we can see 1 class contains more records and 2 class contains few records so it has imbalance problem. In Figure 13, we are calculating skewness of each attribute where negative value refers to unimportant features and positive values refers to important features.

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

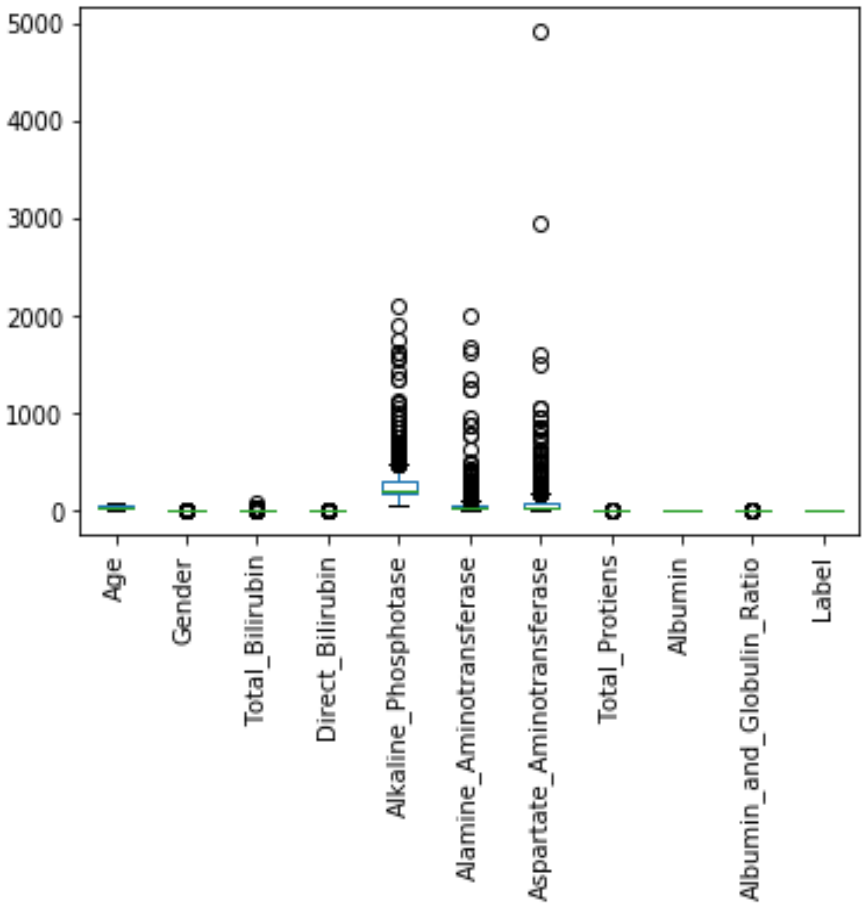


Figure 14. Graphical representation of skewness values.

Table 1. Performance comparison

Method	Accuracy	Precision	Recall	F1-Score
Existing RF without data balancing	65.81	57.99	55.56	55.17
Existing RF with SMOTE balancing	80.83	83.87	82.36	80.75
Existing RF with SMOTE-ENC balancing	76.64	76.67	76.95	76.59
Existing RF with SMOTE-Tomek balancing	84.5	85.73	84.85	84.80
Existing K-means-SMOTE balancing	76.64	77.65	77.65	76.64
Existing SVM with SMOTE balancing	80.23	80.24	79.80	79.95
Proposed RF with SMOTE-ENN balancing	91.39	93.84	88.88	90.47

5. CONCLUSION

The machine learning algorithms presented in this study can support medical experts but are not the alternative when making decisions from ML classifiers for diagnostic pathways. These methods can reduce many of the limitations that occur in healthcare associated with inaccuracy in diagnoses, missing data, cost, and time. Application of the RF-SMOTE-ENN methods can help reduce the total burden of liver disease on public health worldwide by improving recognition of risk factors and diagnostic variables. More importantly, for chronic liver disease, detecting liver disease at earlier stages or in

hidden cases by RF-SMOTE-ENN could decrease liver-related mortality, transplants, and/or hospitalizations. Early detection improves prognosis, since treatment can be given before progression of the disease to later stages. Invasive tests, such as biopsy, would occur less in this case as well. Although this study focused on hepatitis and chronic liver disease variables for ML training, it can be hypothesized that the methods can be used to distinguish other types of liver disease from healthy individuals. Applying all of the mentioned methods to other areas of medicine could open the doors for AI/ML-facilitated diagnosis.

6. FUTURE SCOPE

In the future, the local interpretable model-agnostic explanation method will be used to understand the model's interpretability. Instead of binary classification, one may use multinomial classification by separating the types of liver disease. In this way, each model's performance can be compared. The described ML methods can assist health sectors to achieve a better diagnosis providing effective results in identifying groups or levels within medical data to facilitate healthcare workers. Moreover, ML methods are data driven, and they directly use diagnostic variables from patients' medical tests. Thus, it is a more reliable process. The applied ML methods in this article can save time, costs, and potentially lives for the betterment of disease diagnosis.

REFERENCES

[1] Rong-Ho Lin. An intelligent model for liver disease diagnosis. Artificial Intelligence in Medicine 2009;47:53—62.
[2] Schiff's Diseases of the Liver, 10th Edition Copyright ©2007 Lippincott Williams & Wilkins by Schiff, Eugene R.; Sorrell, Michael F.; Maddrey, Willis C.
[3] 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
[4] Paul R. Harper, A review and comparison of classification algorithms for decision making
[5] Lung-Cheng Huang, Sen- Yen Hsu and Eugene Lin, A comparison of classification methods for predicting Chronic Fatigue Syndrome based on genetic data (2009).
[6] N. A. Shackel, D. Seth, P. S. Haber, M. D. Gorrell and G. W. McCaughan. "The hepatic transcriptome in human liver disease. Comp Hepatol", 2006 Nov 7;5:6. doi: 10.1186/1476-5926-5-6. PMID: 17090326; PMCID: PMC1665460.
[7] R. H. Lin. "An intelligent model for liver disease diagnosis", Artif Intell Med. 2009 Sep;47(1):53-62. doi: 10.1016/j.artmed.2009.05.005. Epub 2009 Jun 21. PMID: 19540738.
[8] B. VenkataRamana, M. Surendra Prasad Babu and N. B. Venkateswarlu, "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis". International Journal of Engineering Reasearch and Development, 2012
[9] Y. Kumar and G. Sahoo. "Prediction of different types of liver diseases using rule based classification model". Technol Health Care. 2013;21(5):417-32. doi: 10.3233/THC-130742. PMID: 23963359.
[10] S. Sontakke, J. Lohokare and R. Dani, "Diagnosis of liver diseases using machine learning,", 2017 International Conference on Emerging Trends & Innovation in ICT (ICEI), 2017, pp. 129-133, doi: 10.1109/ETIICT.2017.7977023.
[11] V. J. Gogi and V. M.N., "Prognosis of Liver Disease: Using Machine Learning Algorithms", 2018 International Conference on Recent Innovations in Electrical, Electronics &

Communication Engineering (ICRIEECE), 2018, pp. 875-879, doi: 10.1109/ICRIEECE44171.2018.9008482.

[12] J. Javad Hassannataj, S. Hamid, D. Abdollah and S. Shahaboddin, "Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection", Informatics in Medicine Unlocked, vol. 17, 2019, 100255, no. 2352-9148, https://doi.org/10.1016/j.imu.2019.100255.

[13] S. Ambesange, V. A, R. Uppin, S. Patil and V. Patil, "Optimizing Liver disease prediction with Random Forest by various Data balancing Techniques", 2020 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), 2020, pp. 98-102, doi: 10.1109/CCEM50674.2020.00030.

[14] M. A. Kuzhippallil, C. Joseph and A. Kannan, "Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients", 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 778-782, doi: 10.1109/ICACCS48705.2020.9074368.

[15] S. Deshmukh, A. Lokhande, R. Wasnik and N. Singhal. "Vacuole Segmentation and Quantification in Liver Images of Wistar Rat", Annu Int Conf IEEE Eng Med Biol Soc. 2020 Jul;2020:1396-1399. doi: 10.1109/EMBC44109.2020.9176500. PMID: 33018250.