Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2704 – 2716
Received 05 March 2021; Accepted 01 April 2021.

Early Forecasting of Chronic Liver Disease from Liver Function Test Imbalance Datasets

Manjunath Varchagall[1],D Sivakumar[2], C SureshKumar[3],Ranjitha V[4], Priyanka N[5]
[1]Research Scholor, Rajrarajeswari College of Engineering, Bangalore, Assistant Professor,Dept of CSE, Gitam School of Technology, Bangalore, Karnataka, India
[2]Rajrarajeswari College of Engineering, Bangalore, Karnataka, India
[3]KGiSL institute of Technology, Coimbatore, Tamilnadu, India
[4,5]Gitam School of Technology, Bangalore, Karnataka, India

mvarchag@gitam.edu[1],dskumarcse@yahoo.co.in[2],cskinit@gmail.com[3], rvenkate@gitam.edu[4], pnarayan@gitam.edu[5]

ABSTRACT

The primary goal of this study is to develop a model for predicting chronic liver disease in its early stages from datasets containing Liver Function Test (LFT) imbalance results, which will aid practitioners in accurately diagnosing liver disease. Detecting disease in its early stages can be difficult, as practitioners often struggle to predict the disease due to its ambiguous symptoms. A total of two data sets were used in this analysis,the second dataset (Primary) was obtained from the Karnataka region of India, and the first dataset (secondary) was taken from the UCI repository. To balance the datasets, we used the Random Forest and K-Nearest Neighbour's (KNN) algorithms, as well as the Synthetic Minority Oversampling Technique (SMOTE). On both the imbalanced and balanced datasets, as well as the various parameters, we compared the effects of the two algorithms. Random forest outperforms KNN in terms of accuracy, specificity, precision, and false positive rate (FPR) on balanced datasets, while KNN outperforms Random forest in terms of accuracy, specificity, sensitivity, FPR, and FNR parameters.On the majority of parameters, the proposed system is expected to increase the balance dataset's performance. The suggested system is as follows: the balance dataset provides a stronger result for the majority of the parameters. The proposed approach aids physicians in correctly diagnosing liver disease at an early stage.

Keywords: Liver Function Test (LFT), Random Forest, SMOTE, K-Nearest Neighbour (KNN).

Introduction

The liver is the human body's main internal organ, where breakdown of red cells of blood in metabolism, fat absorption and detoxification of toxic substances etc. Liver disease, bacteria or viruses, and alcohol intake are all caused by inflammation or compromised hepatocytes infected with fungi. Every year, approximately one million new liver cancer patients are diagnosed worldwide [1], with China ranking second. In 2015, 217,974 (2.44 percent of all deaths) people died from liver disease in India, while 268,857 died from other causes (2.96 percent of total deaths) individuals died in 2019, according to World Health Organization (WHO) statistics. In addition, the death ratio is growing annually and has become India's 10 most common cause of death [2-3]. The liver disease symptoms are difficult to detect early on since many individuals suffer from liver damage but feel healthy [4].

In the modern healthcare industry, Machine learning and data mining methods in use for prediction of diseases from medical datasets. These methods collect valuable insights from the information repository. The correct analysis of the patient in healthcare communities is very much challenging as the symptom of such illnesses is not easy to determine at a prior stage.There is a test called the liver function test (LFT) that can analyse liver disorder prior to symptoms start.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2704 – 2716
Received 05 March 2021; Accepted 01 April 2021.

A significant classification challenge is the diagnosis of liver disorders. Many medical data sets, such as breast cancer, diabetes and liver diseases, suffer from the issue of class imbalance.

The word imbalance means that one class with more or less than the other class is responsible for the number of observations. The conventional classification algorithm doesn't really work well on datasets of imbalances because they consider that all groups in training data have the same number of samples. The minority class data is classified as the majority class in the imbalance dataset problem. The dataset contains four common methods for dealing with imbalance problems (i) Ensemble methods (ii) Sampling Methods (iii) Alteration in traditional classification algorithms (iv) Cost-sensitive methods. To overcome the class imbalance, data balancing is carried out by either under-sampling or over sampling. The method of under-sampling excludes the event from the majority class. Information loss is the downside associated with the under-sampling process, whereas the benefit is that the training period of the model is minimized due to the deletion of data from huge class [5]. The over-sampling technique adds the minority class to the new or duplicate case. The disadvantage of oversampling is that it needs more preparation time. Ensemble methods [6] use a set of classifiers to identify a new unknown instance by weighing their predictions in a weighted vote. It yields better results than a single learner, but the model takes more storage space and training time. The cost-sensitive method of learning correlates the cost with cases that are misclassified.

The cost-sensitive [7] method uses different cost matrices to solve the imbalanced problem, reflecting the costs of wrongly classifying any specific case. The cost-sensitive method of learning needs to identify the cost of misclassification, which is not possible with data sets [8].
Traditional classification procedures are improved to deal with data imbalance problems since the normal distribution of data is not disrupted by this approach. This technique aims to change the conventional algorithms of classification to bias the learning against the minority class. By obtaining the class of their K nearest neighbours that Cover and Hart introduced in 1967, the KNN algorithm identifies the label of class of unclassified points [9].

The primary objective of our research is the early prediction of chronic liver disease from the LFT imbalanced dataset. Two datasets are used for this study one is Indian Liver Patient Dataset (ILPD) collected from UCI repository [8], having 583 patient’s data and other is collected from Karnataka (India) region, having 7865 patients’ records. Classification model does not learn properly when the size of the dataset is small. The datasets that are available on the internet are small in size so we have collected the dataset.

Literature Review

Many studies have been performed in recent years to predict liver disease using global classification methods. The literature has been examined from various perspectives.

Yuantingyan , Ruiqingliu et.al[10], the Proposed research uses CCA and SMOTE techniques and Twenty five imbalance dataset are considered for their analysis. The outcome of the research was to provide cleaning the balanced dataset which improved the efficiency of SMOTE. The major limitation is that longer training time is needed because of oversampling.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2704 – 2716
Received 05 March 2021; Accepted 01 April 2021.

Somaya Hashem1, Gamal Esma et.al [11], The proposed work suggested that to use ALT (Alternative decision tree) GA, multi-linear regression and PSO techniques. The National Viral Hepatitis Monitoring Committee, database, Egypt dataset is considered for their study. The highest accuracy of 84.4 percent was provided by established models for advanced fibrosis risk factors and ADT techniques. ADT offers greater accuracy but minimum sensitivity (7%).

Sujit Kumar1·Saroj Kr. Biswas et.al [12], The proposed research uses the TLUSBoost(Tomek-link based undersampling and boosting technique).The Sixteen Datasets Imbalanced are considered for their analysis. The model shows the LUSBoost model is superior to most parameters in the BalanceCascade, EasyEnsemble, SMOTEBoost and RUSBoost methods.

Kwabena EboBennin and Jacky Keung [13], the proposed work uses NNET, MAHAKIL, C4.5, RF, SVM, KNN classifier with ROS, SMOTE sampling techniques, Borderline-SMOTE, and ADASYN. The Twenty Datasets Imbalanced is considered for their research study. The study's findings indicate that a new over-sampling approach has been proposed for detecting programme vulnerabilities, which implements instances based on their distance from Mahalanobis and tackles SMOTE's redundant data generation challenges. The key drawback is the ban on oversampling, which necessitates further preparation time.

Harshita Patel and Ghanshyam Singh Thaku [14], The proposed research demonstrated the use of Fuzzy-NWKNN (Hybrid fuzzy weighted nearest Neighbor) and it works on six imbalance datasets. The research outcome shows that the Fuzzy-NWKNN approach is a next level of the NWKNN technique. The limitation of the research is assigning a weight for data of big and small groups, subsequently in certain circumstances weight calculations fail.

JoonhoGonga and HyunjoongKim[15], The proposed work that uses RHSBoost(Random hybrid sampling boosting Algorithm) and 16 datasets for imbalance are chosen for their analysis. The outcome of the proposed work is that RHSBoost is a mixture of hybrid sampling and AdaBoost. RHSBoost offers good classification output over different imbalanced datasets. The limitation of the sampling approach is related to the RHSBoost

MoloudAbdar, Mariam Zomorodi-Moghada et.al [16], The proposed work suggested that to use CHAID and Boosted C5.0 and also Indian Liver Patient Dataset (ILPD) is considered for their study. The outcome of the research is that 93.75% of accuracy was generated by Boosted C5.0, while 65.00% of accuracy was generated by CHAID algorithms. The major research drawback is operated on a single collected dataset from the UCI repository.

Neil Yuwen Ye et.al[17], The proposed research uses the C5.0, CART, CHAID with MLPNN and boosting technique and also Indian liver Patient Dataset is considered for their study. The outcome of the results are Operated on a single collected dataset from the UCI repository. Overall 94.12%, MLPNNB-C5.0 produces better accuracy.

Dan Meng, Libo Zhang et.al[18], The proposed work uses the Fully connected network (FCNet) and Transfer learning (TL) methods and also 279 Ultrasound images with ROI is considered for their analysis. The outcome of the research is a system for the categorising of liver fibrosis on ultrasound images has been suggested.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2704 – 2716
Received 05 March 2021; Accepted 01 April 2021.

Qi Kang, XiaoShuang Chen et.al[19], The research work uses the Adaboost, Noise-Filtered Under-Sampling,UnderBagging, RUSBoost, and Easy Ensemble. The 16 datasets for imbalance are considered for their analysis. The outcome of the proposed results is A novel under-sampling technique has been developed by using an impurities filtering that deletes the outlier from the minority class. The limitation Data losses are a side effect of the under-sampling technique.

Xiaofeng Zhou, Youglai Zhang et.al[20], The proposed work uses GSO with SVDD algorithm and Collected community LFT data from Beijing hospital for their study. The proposed approach generates 84.28 percent of the Accuracy, 96 percent sensitivity, and 96 percent sensitivity Specificity of 86.28 percent. A procedure is applied to a sample Records of 225 patients from 1000 Liver Function Research Patients about data records.

Lizhipeng, Hongli Zhang et.al[21], The proposed research work uses Standard Algorithms: KNN, C4.5, RandomForest, PNN, SVM Imbalanced Algorithms: AdaBoost, SMOTEBagging, SMOTEBoost, Bagging also uses six datasets for traffic imbalances and is chosen for their analysis. The outcome of the research is Established imbalanced gravitation of data Classification based upon classification (IDGC) method of overcoming the Internet imbalance issues with traffic recognition. The limitation is that just half the data sets, the proposed IDGC obtained the maximum AUC values.

Chris Seiffert, Taghi M. Khoshgoftaar et.al[22], The proposed work uses a method for RUSBoost that combines the boosting process and the random method undersampling and also uses 15 imbalance datasets for their analysis. Comparison of the findings of the RUSBoost process together with AdaBoost, RUS, SMOTE and SMOTEBoost and discovered that RUSBoost creates beneficial outcomes. The limitation of the research is the technique of boosting strengthens the Classifier complexity.

Rong Ho Lin and Chun Ling Chaung[23], the proposed research demonstrated that uses AHP), Artificial neural network (ANN), and case-based reasoning and also Health Review information for Taiwan Medical Center dataset are chosen for their study. The outcome of the research is Build an approach to ILDM for Liver disease prediction. The disadvantage of the proposed research is the CBR method requires greater storage space for all cases and more time in processing in order to find a related case library case.

Songbo Tan[24], the proposed work uses Classification of text document NWKNN imbalanced and also Reuter and the document corpus of TDT2 is considered for their study. The proposed outcome proved that NWKNN performed well when compared to KNN. The limitation is that NWKNN works for the automated classification of text documents with imbalances.

Methods Used

a) Random Forest Tree Classification

Name suggests that it consists of collecting single decision trees that act as an ensemble. Each tree spits out in a random forest, forming a class with the others. highest votes have become the identification of the model we are considering is depicted in the diagram below. Performance of

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2704 – 2716
Received 05 March 2021; Accepted 01 April 2021.

random-forest is very well in data science because many reasonably uncorrelated prototypes (trees) acting as a committee would better all of the individual models' behavior.

The secret is the low correlation among models. If low-correlation statements taken together to construct a more extensive portfolio compared to the sum of models, uncorrelated models may establish ensemble results that are more important compared to any of the predicted outputs. The justification for these beautiful facts is that the trees shield each other from their mistakes. Although some trees will be wrong, many other trees are suitable so that the trees which are correct need to be considered for a move. Therefore, the basic requirements for sound output of random forests are: There must be some actual signal throughout our parameters so that models created using these features do better than usual guessing. The predictions produced by the trees individually required to have minimum correlations with others. The below fig1 demonstrate the working of Random Forest tree techniques

Algorithm Random forest tree

Step 1: The collection of samples randomly from a given dataset begins.
Step 2: This algorithm is basedon creating a decision tree. Hence, from any decision tree, prediction results can be considered.
Step 3: For any expected outcome, voting will be carried out.
Step 4: At the end pick the outcome of the highest voted prediction as the final result of the prediction

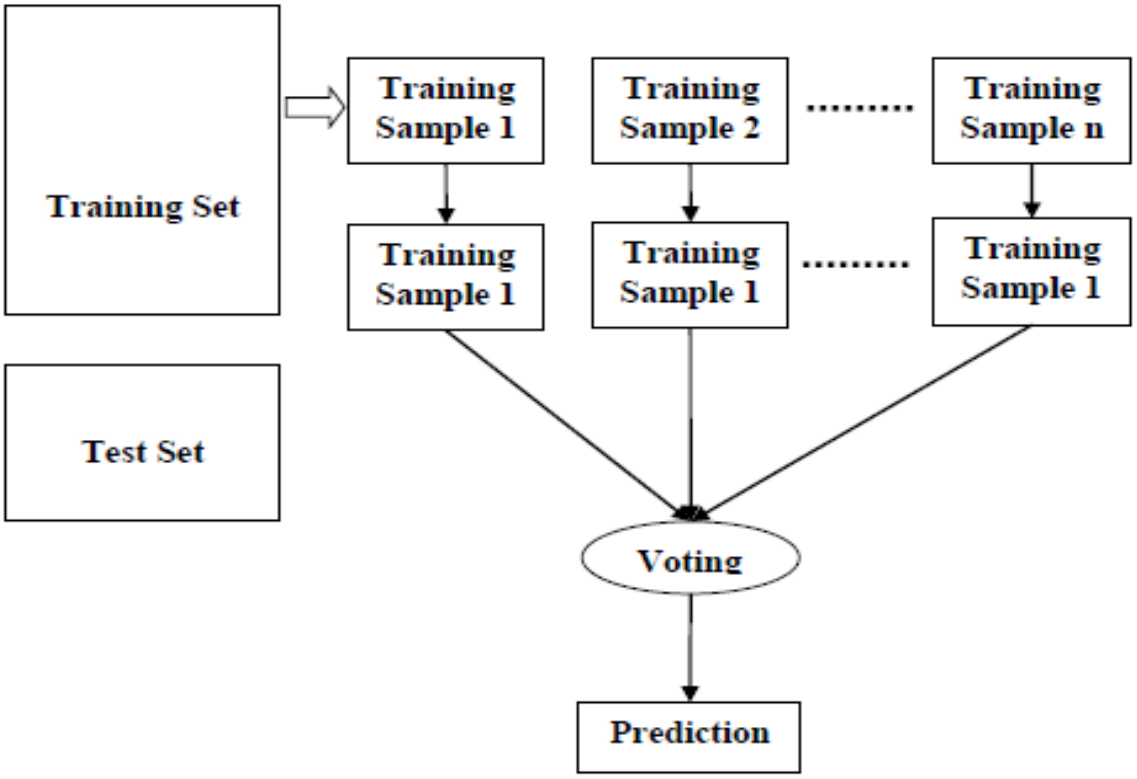


Fig1. Working of Random Forest Tree Classification

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2704 – 2716
Received 05 March 2021; Accepted 01 April 2021.

b) K-NN (K- Nearest Neighbor)

K- Nearest neighbor classifier identifies the class of an unknown instance by obtaining the K-nearest neighbor's class. The new instance will be labeled with the class of the highest frequency form the K most similar instances [22, 23]. The algorithm is work as follows:

1. Let X = (X_1, y_1), (X_2, y_2), …… (Xn, yn), where Xi is data points, xi € Rd, yi is labeled class corresponding to X_i and let yi € (+1, -1)
2. Find the D (X_{new}, Xi) where: X_{new} is the instance which class is to find and D is the distance function which finds the distance between X_{new} to Xi.
3. Arrange the distance in ascending order
4. Take first K sorted distances from list
5. Assign highest frequency class of first K sorted distance data point to the X_{new}

c) SMOTE (Synthetic Minority Oversampling Technique)

The Class imbalance problem is a significant issue faced by medical datasets. The datasets are imbalanced if one of the classes contains fewer instances than the other categories. Classifiers produce a combined prediction for minority classes. Synthetic Minority Oversampling Technique (SMOTE) oversamples the minority class, where new synthetic observations are generated. By the SMOTE algorithm, synthetic statements, creations based on feature space similarities between existing minorities instances can be defined as

D: 'dimensional dataset'

S_mĆ D: S_m is the minority class instance.

χ_iЄ S_m: χ_i is the minority class instance under consideration

δ Є rand(0,1)

χ_{sync}: Synthetic observation or instance

The Synthetic Minority Oversampling Technique typically requires the following key steps:

Step1: $\forall xi$, find K nearest neighbor in the feature space

Step2: Randomly select one of the neighbor of x_i called $x\hat{}_i$

Step3: Take the difference between and x_i called $x\hat{}_i$

Step4: Multiply the difference with δ

Step5: Find the new point or observation χsync on the line segment by adding the obtained value to the feature vector x_i

Step6:$\forall xi$, repeat the step 2 to step 5

All these steps can be represented simply by the following equation

$$Xsync = xi + (x\hat{}i - xi) * \delta \quad (1)$$

d) DATASETS

In this work, two liver patient datasets used to build and test the models. The Karnataka Area Liver Patient Dataset is the first dataset obtained from the Indian state of Karnataka (KRLPD). This dataset includes 12 important Liver Function Test (LFT) features age, sex, a/g ratio, albumin, alkaline Phosphatase, direct bilirubin, globulin, indirect bilirubin, sgot, sgpt, total bilirubin, total protein, having two classes. This dataset consists of 7865 records, in which 6282 persons have liver disorder and 1583 persons having health.

Table1. Describes the attributes in KRLPD

Sl no	Name of Attribute	Attribute Type	Ranges
1	Age	Interval	6-92
2	Gender	Nominal	Male-Female

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2704 – 2716
Received 05 March 2021; Accepted 01 April 2021.

3	A/G Ratio (Albumin Globulin Ratio)	Interval	0.21-2.04
4	ALB(Albumin)	Interval	1-4.90
5	Alkphos (Alk. Phosphatase)	Interval	37-1591
6	DB(Direct Bilirubin)	Interval	0-40.21
7	Globulin	Interval	2-9
8	Indirect Bilirubin (IB)	Interval	0.10-14.91
9	SGOT	Interval	13-1360
10	SGPT	Interval	16-2232
11	Total Bilirubin (TB)	Interval	0.20-55.11
12	Total Protein (TP)	Interval	3.61-10.23
13	Predictor	Binary	0-1

The dataset also contains information about 5056 males and 2809 females. The description of attributes of KRLPD is shown in Table 1. Another one is the Indian Liver Patient Dataset (ILPD) collected from UCI repository [8], having 11 essential LFT features. This dataset consists of 583 patient records in which 416 persons have the liver disorder and 167 healthy persons.

Proposed Approach

This paper has used the Random forest tree and K-Nearest Neighbour (KNN) algorithms with or without SMOTE to find the liver disorder form the imbalanced Liver Function Test dataset. In this regard, we have used MATLAB R2014 to evaluate the result on KRLPD and ILPD datasets. Fig.2 shows all the steps of implemented work clearly.

Implementation of this work follows the following steps-

1. Two datasets have been used in this study. The first ILPD dataset was chosen and preprocessed from the UCI repository [8]. The second KRLPD dataset has been gathered and preprocessed.
2. Perform data balancing of ILPD and KRLPD imbalanced datasets using Synthetic Minority Over-sampling Technique.
3. Train the model using Random Forest/KNN (on four folds) for both balanced and imbalanced dataset of ILPD/KRLPD.
4. Predict the label on the rest of one test set using the SVM/KNN train model.
5. Find average results on different parameters of independent test sets and compare the product of an imbalanced and balanced dataset of ILPD/KRLPD

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2704 – 2716
Received 05 March 2021; Accepted 01 April 2021.

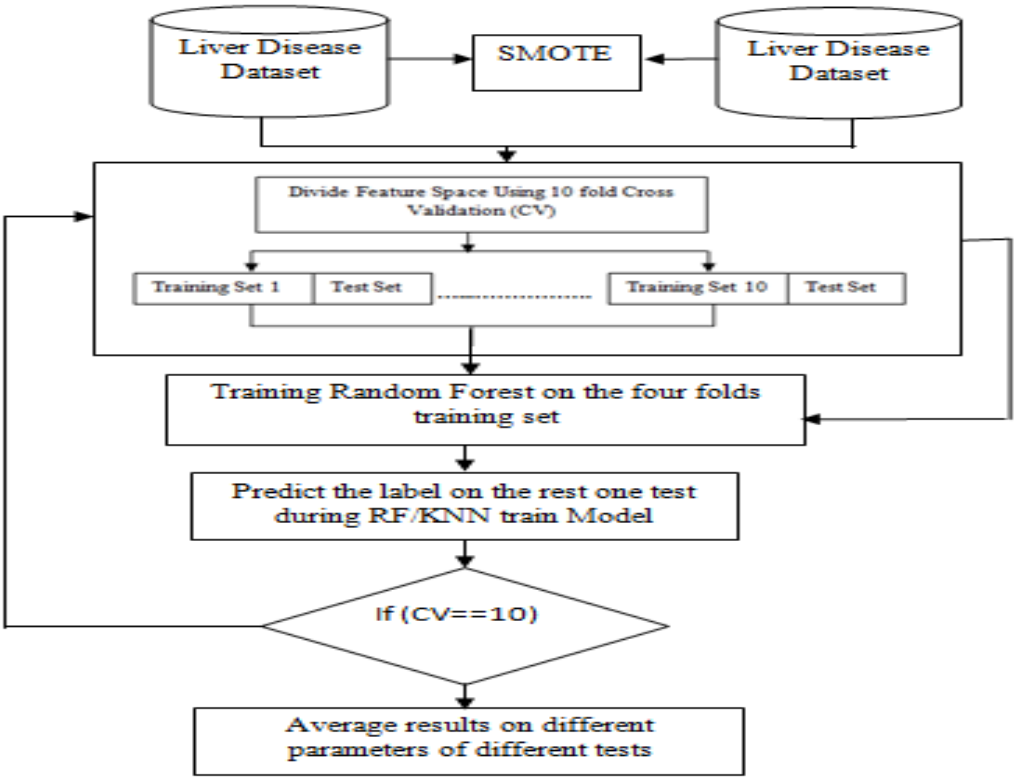


Fig2. Flow Diagram of Proposed Approach

Performance Measures

Performance metrics are used to access the classification models. These metrics are confusion matrix, sensitivity, specificity, precision, False Negative Rate (FNR), False Positive Rate (FPR), and accuracy of classification.

Confusion Matrix: A classification matrix real and expected effects are summarized in the uncertainty matrix [29]. As shown in Table 2, the uncertainty matrix distinguishes the number of right and incorrect predictions with count values and divides them into groups [30].

Table2: Describes the attributes in KRLPD

Data Class	True classification	False Classification
True	Correct(TP)	Incorrect(FP)
False	Incorrect(FN)	Correct(TN)

Accuracy

The accuracy of a classifier is the number of correct predictions from all predictions made. If the dataset is imbalanced then accuracy alone may not justify the model.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Specificity (TNR)

The accuracy of the data that is classifiedin the negative class.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2704 – 2716
Received 05 March 2021; Accepted 01 April 2021.

$$Specificity = \frac{TP}{TP+FP} \quad (3)$$

False Positive Rate (FPR)
Percentage of miss classified (Error) in Negative Class

$$False\ Positive\ rate = \frac{FP}{FP+TN} \quad (4)$$

False Negative Rate (FNR)
Percentage of miss classified (Error) in positive class.

$$False\ Negative\ rate = \frac{FN}{FN+TP} \quad (5)$$

Results and Discussion

In this study, Random Forest and KNN (for k=3) algorithms and oversampling technique is applied. 10-fold cross-validation is used for generating unbiased outcome. The detailed outcomes of the performance matrices are shown in Tables 3 and Table 4. According to Table 3, it can be seen that Random forest on KRLPD gives the better result for the parameter accuracy 97.42%, specificity 95.39%, precision 92.12%, and FPR 5.61% with the balanced dataset, whereas sensitivity 96.53% and FNR 1.47% with the imbalanced dataset.

Table 3 also shows that the Random Forest on ILPD produce the best result for the accuracy 74.96%, specificity 71.59%, precision 66.15%, FPR 28.41% and FNR 3.35% with the balanced dataset, whereas sensitivity 85.14% and FNR 12.86% with the imbalanced dataset.

Table3: Performance Measure using Random Forest Tree

Performance metrics	Random Forest on KRLPD		Random Forest on ILPD	
	Imbalanced dataset	Balanced Dataset	Imbalanced dataset	Balanced Dataset
Accuracy	91.38	97.42	66.21	74.96
Specificity	72.09	95.39	42.63	71.59
Sensitivity	96.53	95.65	85.14	77.98
Precision	91.9	95.12	62.33	66.15
FPR	25.91	5.61	55.37	28.41
FNR	1.47	3.35	12.86	3.35

Table3: Performance Measure using KNN

Performance metrics	KNN on KRLPD		KNN on ILPD	
	Imbalanced dataset	Balanced Dataset	Imbalanced dataset	Balanced Dataset
Accuracy	76.16	82.42	65.13	75.96

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2704 – 2716
Received 05 March 2021; Accepted 01 April 2021.

Specificity	37.66	75.72	38.04	71.44
Sensitivity	84.51	93.11	72.79	82.43
Precision	86.43	65.85	78.34	63.15
FPR	60.34	24.28	61.96	28.35
FNR	15.49	5.89	25.21	17.57

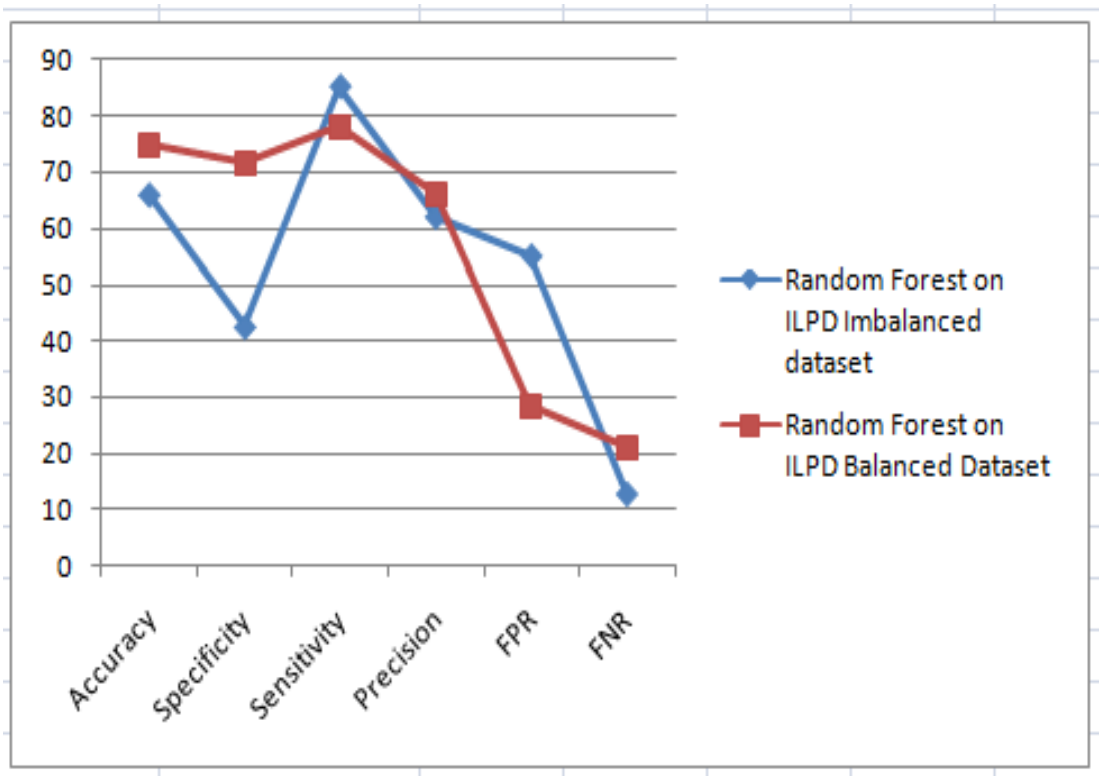
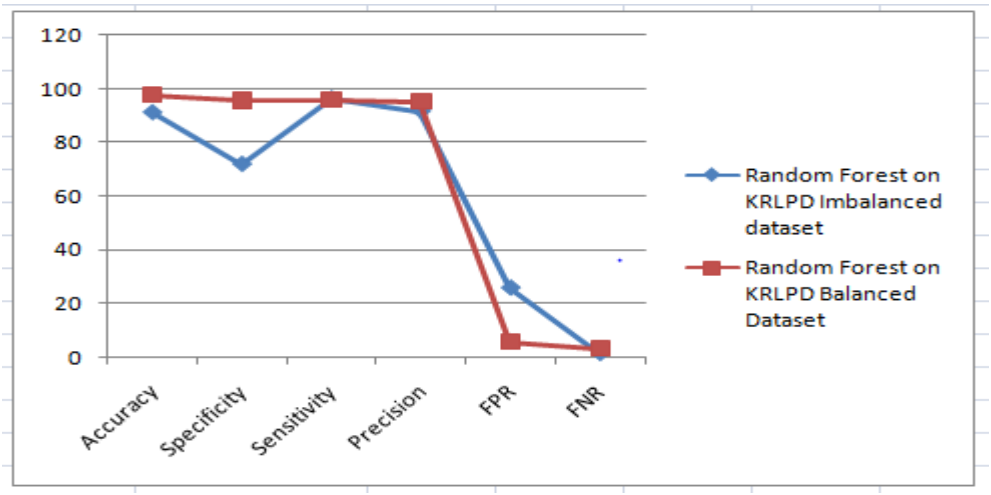


Fig3. A comparison between the performance of imbalance and balance dataset of ILPD using Random forest on different parameters

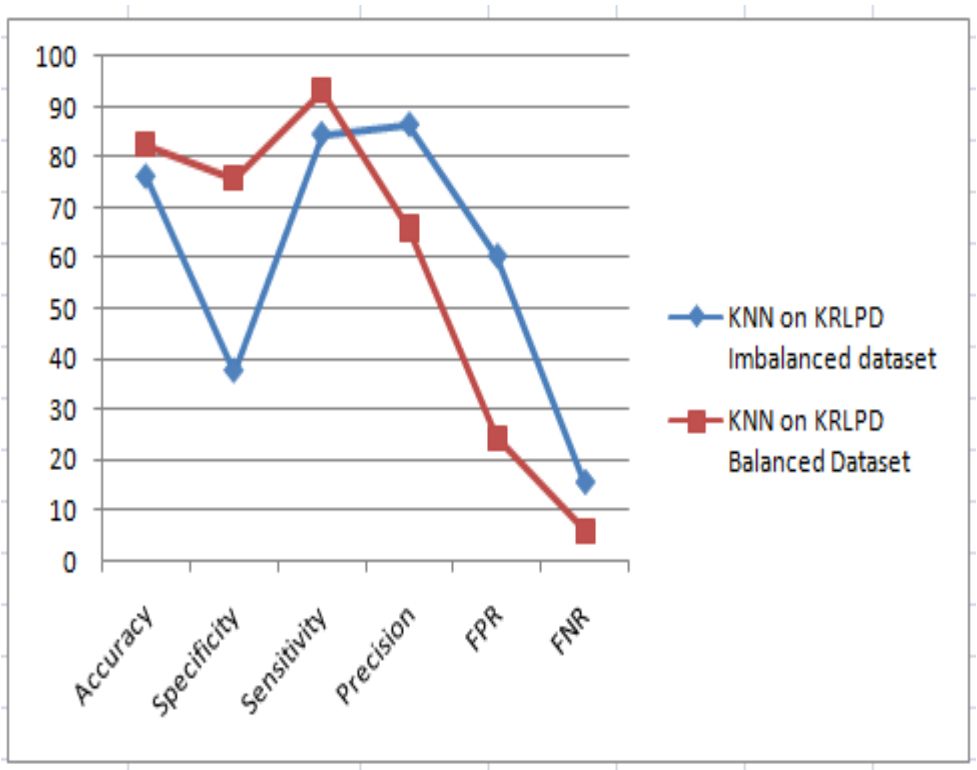
Fig. 3 and Fig. 4 show a comparative graphical representation of performance for imbalance and balance dataset using Random forest. In graphical representation horizontal axis represents Performance metrics and Vertical axis represents measured value in (%). However, according to Table 4, we found that KNN on KRLPD produced the better result for accuracy, specificity, sensitivity, FPR and FNR were 81.42%, 74.72%, 93.11%, 25.28% and 6.89% respectively for the balanced dataset whereas precision 87.43% for the imbalanced dataset. Table 4 also shows that the KNN on ILPD produce the best result for accuracy, specificity, sensitivity, FPR and FNR were 74.67%, 70.44, 81.43, 29.56 and 18.57 respectively for the balanced dataset whereas precision 79.34 % with the imbalanced dataset. Fig. 4 and Fig. 5 show a comparative graphical representation of performance for imbalance and balance dataset using KNN.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2704 – 2716
Received 05 March 2021; Accepted 01 April 2021.



Performance metrics

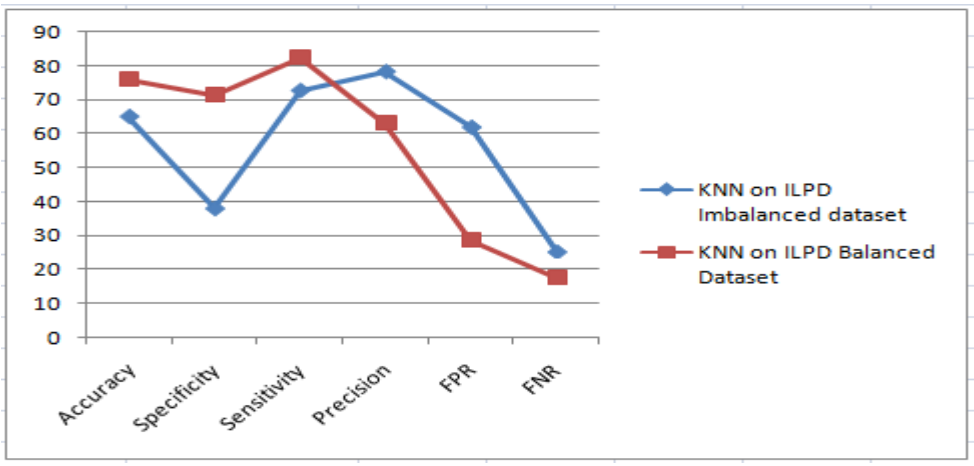
Fig4. A comparison between the performance of imbalance and balance dataset of KRLPD using Random forest on different parameters



Performance metrics

Fig5. A comparison between the performance of imbalance and balance dataset of ILPD using KNN on different parameters

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 2704 – 2716
Received 05 March 2021; Accepted 01 April 2021.



Performance metrics

Fig6. A comparison between the performance of imbalance and balance dataset of KRLPD using KNN on different parameters

Conclusion and Future Enhancement

The imbalanced liver function research dataset was used in this analysis to predict liver disease. Imbalanced datasets often have poor accuracy in all of the dataset's groups. For dataset balancing, we used a synthetic minority oversampling technique.Two well-known algorithms, Random forest and KNN, were used on both the imbalance and balance datasets of the ILPD and KRLPD in this regard. On a balanced dataset with most of the parameters, our proposed system produces a better performance. For further enhancement the life quality attributes to be considered for analysis.

References

1. Chuang, C.-L. (2011). Case-based reasoning support for liver disease diagnosis. Artificial Intelligence in Medicine, 53(1), 15–23. doi:10.1016/j.artmed.2011.06.002
2. Lin, R.-H. (2009). An intelligent model for liver disease diagnosis. Artificial Intelligence in Medicine, 47(1), 53–62. doi:10.1016/j.artmed.2009.05.005.
3. Khakhar A (2017) A liver disease in india. http://www.livertransplant.org/liver-transplantation/ awareness/liver-diseases-in-india-stats, accessed on 08/04/2019
4. Media L (2017) World health ranking. https://www.worldlifeexpectancy.com/india-liver-disease, accessed on 08/04/2019.
5. Alfisahrin SNN, Mantoro T (2013) Data mining techniques for optimization of liver disease classification. In: 2013 International conference on advanced computer science applications and technologies.IEEE, pp 379–384.
6. TAN, S. (2005). Neighbor-weighted K-nearest neighbor for unbalanced text corpus. Expert Systems with Applications, 28(4), 667–671. doi:10.1016/j.eswa.2004.12.023.
7. Lin, R.-H., & Chuang, C.-L. (2010). A hybrid diagnosis model for determining the types of the liver disease. Computers in Biology and Medicine, 40(7), 665–670.

8. Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. IEEE.
9. Peng, L., Zhang, H., Yang, B., & Chen, Y. (2014). A new approach for imbalanced data classification based on data gravitation. Information Sciences, 288, 347–373.
10. Zhou, X., Zhang, Y., Shi, M., Shi, H., & Zheng, Z. (2014). Early detection of liver disease using data visualisation and classification method. Biomedical Signal Processing and Control, 11, 27–35. doi:10.1016/j.bspc.2014.02.006.
11. Kang, Q., Chen, X., Li, S., & Zhou, M. (2017). A Noise-Filtered Under-Sampling Schemefor Imbalanced Classification. IEEE Transactions on Cybernetics, 47(12), 4263–4274.
doi:10.1109/tcyb.2016.2606104
12. Meng, D., Zhang, L., Cao, G., Cao, W., Zhang, G., & Hu, B. (2017). Liver fibrosis classification based on transfer learning and FCNet for ultrasound images. IEEE Access, 1–1. doi:10.1109/access.2017.2689058
13. Abdar, M., Yen, N. Y., & Hung, J. C.-S. (2017). Improving the Diagnosis of Liver Disease Using Multilayer Perceptron Neural Network and Boosted Decision
14. Abdar, M., Zomorodi-Moghadam, M., Das, R., & Ting, I.-H. (2017). Performance analysis of classification algorithms on early detection of liver disease. Expert Systems with Applications, 67, 239–251. doi:10.1016/j.eswa.2016.08.065.
15. Gong, J., & Kim, H. (2017). RHSBoost: Improving classification performance in imbalance data. Computational Statistics & Data Analysis, 111, 1–13.
16. Patel1 Harshita, Thakur Ghanshyam Singh: Classification of Imbalanced Data Using a Modified Fuzzy-Neighbor Weighted Approach, International Journal of Intelligent Engineering and Systems, Vol.10, No.1, 2017.
17. Bennin, K. E., Keung, J., Phannachitta, P., Monden, A., & Mensah, S. (2018). MAHAKIL: Diversity Based Oversampling Approach to Alleviate the Class Imbalance Issue in Software Defect Prediction. IEEE Transactions on Software Engineering
18. Umar, S., Biswas, S. K., & Devi, D. (2018). TLUSBoost algorithm: a boosting solution for class imbalance problems. Soft Computing. doi:10.1007/s00500-018-3629-4
19. Hashem, S., Esmat, G., Elakel, W., Habashy, S., Raouf, S. A., Elhefnawi, M., ElHefnawi, M. (2018). Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients. IEEE/ACM Transactions on Computational Biology and Bioinformatics.
20. Yan, Y., Liu, R., Ding, Z., Du, X., Chen, J., & Zhang, Y. (2019). A Parameter-free Cleaning Method for SMOTE in Imbalanced Classification. IEEE Access, 1–1doi:10.1109/access2019.2899467.