Early Detection of the Liver Disorder from Imbalance Liver Function Test Datasets

Pushpendra Kumar, Ramjeevan Singh Thakur

Abstract*: Aim of this research is to develop a model for early detection of liver disorder from imbalance Liver Function Test (LFT) results' datasets that assists the practitioners in diagnosing the liver disease efficiently. Because in the initial stage symptoms of the diseases are vague so the medical practitioners often fail to detect the disease. This study used two datasets of Liver Function Test (LFT) for building the systems, one is ILPD dataset (secondary) taken from UCI repository and second dataset (Primary) is collected form Madhya Pradesh region of India. We have used Support Vector Machine and K-Nearest Neighbour (KNN) algorithms to implement the system and Synthetic Minority Oversampling Technique (SMOTE) to balance the datasets. We have compared the results of both the algorithm on the different parameter for both the imbalanced and balanced datasets. We get the improved result for accuracy, specificity, precision, false positive rate (FPR) parameters on balanced datasets using SVM whereas using KNN we get improve results for accuracy, specificity, sensitivity, FPR and FNR parameters on balanced datasets. We can conclude that the proposed system gives the improve result on balance dataset on most of the parameter. Proposed system helps the healthcare practitioners in diagnosing the liver disease efficiently at the early stage.

Index Terms*: K Nearest Neighbor (KNN), Liver Function Test (LFT), SMOTE, Support Vector Machine (SVM).

I. INTRODUCTION

Nowadays the advancement of computer technology has notably improved the data formation in various organizations such as the university, bank, hospital, e-commerce sites, and many others. This rapid growth of data in the various organizations insists us to interpret and extract useful information from them. Data mining is a tool that identifies and extracts the valuable information from the huge amount of data very preciously in less time and cost [1, 2]. Data generated by the medical institute are very important for the knowledge extraction purpose that helps the medical practitioner in fast diagnosing the disease and providing better treatment to patients. The accurate diagnosis of patients and providing proper treatment is very important in medical science. Wrong medication may lead to wastage of money and time for the patients, sometimes this may lead to the irreparable loss (death) [3]. One of the fatal diseases that have affected one in five persons of India is liver disease. It is expected that India may become the "world capital" for liver disease by 2025 [4]. The liver is one of the important and largest organs of the body that is situated in the upper right portion of the stomach and under the diaphragm [2]. The weight of liver is about 1.36 kg and reddish brown in color. The liver performs more than 500 functions, some well-known functions are the production of bile, production of important proteins for blood clotting, purification of blood, helping in fat digestion, decomposing red blood cells and detoxifying harmful chemicals [5]. Cause of liver disease can be excesses consumption of alcohol, obesity, infection with viruses, genetic disorder, contaminated food and etc. Medical practitioners often fail to detect the liver disease at the earlier stage because the symptoms of the disease are vague at the initial stage. You may get the symptoms when your liver already has damaged. General symptoms of liver disease are right upper quadrant abdominal pain, nausea, vomiting, jaundice, weakness, Fatigue and weight loss [6]. Liver disease can be of various types, the most common diseases are Viral Hepatitis, the cause of it can be a virus A, B or C; Fatty Liver Disease(FLD), generally FLD happens to those persons who are suffering from the overweight problem, have metabolic syndrome or diabetes; Autoimmune Liver Disease, in which immune system of the body attacks to the liver; Alcoholic Liver Disease, it can be the cause of excesses use of the alcohol; or Genetic Liver Disease [5-7]. These common liver diseases may lead to Liver Cirrhosis and Liver Cirrhosis may lead to liver cancer or failure.

The aim of our study is to early detection of the liver disorder from imbalance liver function test dataset. Two datasets are used for this study one is Indian Liver Patient Dataset (ILPD) taken from UCI repository [8], having 583 patients records and other is collected from Madhya Pradesh (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.

The rest of paper is organized as follows. In section II Related work is discussed. In section III Materials And Methods are discussed. In section IV Datasets are discussed. In section V Performance Measures are discussed. Section VI describes the proposed approach. In section VII Results and Discussion. Finally, in section VII conclusion is given

II. RELATED WORK

In last few years, numbers of investigations have been carried out on different kinds of diseases by using the data mining techniques around the world. Bahramirad, Mustapha and Maryam Eshraghi (2013)[9] build a classification model and improve their result by performing brute force optimization and Bayesian

Revised Manuscript Received on 8 February 2019.

Pushpendra Kumar, Department of Computer Applications, Maulana Azad National Institute of Technology, Bhopal, 462003, India

Ramjeevan Singh Thakur, Department of Computer Applications, Maulana Azad National Institute of Technology, Bhopal, 462003, India.

Boosting in order to select the best feature. Abdar(2015) [10] Compare the performance of two data mining tool, Rapid Miner and SPSS Modeler on ILPD dataset, is available in UCI repository. Comparing among different algorithm by Rapid Miner tool, SVM algorithm performs better with 72.54% of Accuracy and 100% of Precision and C5.0 algorithm performs better with 87.91% of Accuracy by IBM SPSS Modeler. Alemayehu & Berger (2016) [11] Perform a systematic overview of big data: transforming drug development and health policy decision making and recommend steps to be taken for efficient integration of those. Alfisahrin and Mantoro (2013)[12] perform the comparison of three data mining technique Decision Tree, Naive Bayes, and NBTree on Indian liver disease dataset. Conclude that NBtree have the highest accuracy with 67.01% but slowest computation time having 2.51 seconds and the Naive base has the lowest accuracy with 56.14% but highest computation having 0.04 seconds. Hammad, AbouRizk (2014) [13]Recommend a model based on Knowledge Discovery in Data to analysed and extract useful knowledge from industrial construction project's data. The model understands the existing historical data from developed projects and finds some useful knowledge for the future projects. Data for this model was collected from western-Canadian Structural Steel Fabricator project to demonstrate the ability of the framework. Priya, Juliet and Tamilselvi (2018)[14]Examine the liver patient dataset and build the classification model to predict liver disease. Performance of accuracy of this model is improved by selecting best feature using Particle Swarm Optimization (PSO). J.48 classification algorithm with PSO feature extraction model gives an accuracy of 95.04% while without PSO feature extraction gives 68.77 % accuracy. Abdar,Yen and Hung (2017) [15] Implemented and compare several classification methods using boosting technique in two phases. In the first phase, the researcher has compared B-CHAID, B-CART and B-C5.0 methods and found that accuracy of B-C5.0 is 92.61% better than B-CHAID is 64.77% and B-CART is 67.61%. In the second phase proposed the hybridization of B-CHAID, B-CART and B-C5.0 with Multilayer Perceptron Neural Network (MLPNN), the result shows that accuracy is improved by 1.51%, 14.57% and 12.08% respectively. Hassoon, Kouhi, Moghadam and Abdar (2017) [3]Presents a novel method for liver disease diagnosis. This method optimizes the rules obtain from Boosted C5.0 classification algorithm using the Genetic algorithm, that improves the efficiency and accuracy. Instead of generating 92 rules by Boosted C5.0, GA produces only 24 rules and also accuracy is increased by 12%. Xiaofeng Zhou et al. (2014) [16] Implement a visualization and diagnostic method for liver disease dataset that is collected from a community hospital in Beijing. This method combines the support vector data description (SVDD) with glowworm swarm optimization (GSO) algorithm to enhance the performance of the proposed method, the result shows that implemented method produces 84.28% of accuracy, 96% of sensitivity, and 86.28% of specificity. Moloud Abdar et al. (2017) [2]Compare the performance of two novel classification algorithm, Boosted C5.0 and CHAID on ILPD dataset. The result shows that Boosted C5.0 has an accuracy of 93.75% which is better than 65.00% of accuracy by CHAID. Boosted C5.0 algorithm also considers gender in liver disease prediction and the result shows that females are more susceptible than male. Ramana, Babu and Venkateswarlu (2012) [17] Compare the results of eleven classification methods on ILPD and BUPA liver patient datasets in terms of accuracy, precision, recall. The result shows that logistic method generates the highest 73.39% of accuracy for ILPD dataset whereas the Neural Net and Gaussian Processes generate the highest 73.91% of accuracy for BUPA dataset. Later researcher improves the performance of classification method by selecting best feature using Bayesian Boosting Optimization and Brute Force Optimization. Kant & Ansari (2016) [18]Proposed a method to classify the liver patient dataset using the K-mean algorithm. The result of K-mean is improved by initial seed selection using Atkinson index technique; the improved algorithm produces 23% of accuracy whereas simple K-mean produces 16% of accuracy.

III. MATERIALS AND METHODS

A. Support Vector Machine

Support Vector Machine (SVM) is one of the most popular methods of supervised machine learning algorithm that can be used for classification and regression problems, introduced in 1995 by Cortes and Vapnik [19]. Originally the SVM was developed for classification of linear data in two class, later it was improved that can classify the multi-classes and nonlinear data. It is based on the idea of decision hyper-planes that define the decision boundaries. Decision hyperplane separates the set of the object having a different class[20]. In this algorithm, if we have the N-dimensional dataset (where N is the no. of the feature in a dataset) then we plot each training data points in N dimensioned space. Then we perform classification by dividing the training data points into K (where K is the number of the classes in the dataset) separate regions by hyper-planes of N different dimensions. Later to find the class of the data points, the data points are plotted in the same N-dimensional space, the points are classified into a particular class depending on the region in which the point fall.

The SVM algorithm works as follows for linearly separable dataset

1. Initialize the dataset D as (X_1,y_1), (X_2,y_2), …… (X_n,y_n), where X_i is training tuple, $X_i \in R_d$, y_i is labeled class corresponding to X_i and let $y_i \in (+1, -1)$
2. Corresponding to each training tuple Xi let the weight vector W_i then the optimal hyperplane is computed as

$$W.X + b = 0 \quad (1)$$

Where b is the bias.

3. If any point lies above the separating hyperplane then that point follow the following inequality

$$W.X + b > 0 \quad (2)$$

4. If any point lies below the separating hyperplane then that point follow the following inequality

$$W.X + b < 0 \quad (3)$$

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

5. To define the sides of the margin the weight can be adjusted by computing the following equations

$H_1 : W.X + b \geq 1$ for y_i=+1 (4)

$H_1 : W.X + b \leq 1$ for y_i=-1 (5)

The equation (4) and (5) jointly written as

$y_i(W.X + b) \geq 1 \ \forall i$ (6)

6. The X_i that satisfies the equation (6) is called Support Vector.

7. To maximize the margin between two separating hyperplane compute the following equation

$$Max \frac{2}{\| W \|} \quad \text{(or)} \quad Min \quad \| W \| \qquad (7)$$

Subject to the condition equation (6)

8. By solving the equation (7) using Lagrangian formulation and Karush-Kuhn-Tucker (KKT) conditions we can rewrite the equation as-

$$f(X) = \sum a_i y_i X_i^T X + b \qquad (8)$$

Where: X is the test point, X_i is the support vectors, y_i is the class label of support vector X_i; a_i and b are numeric parameter automatically determine by SVM algorithm

The real world data are not always linearly separable in d-dimensional space, in such case no any linear hyperplane can be formed that would differentiate the classes. There are two methods of resolving the issues with linearly inseparable data. (1) Firstly we map the data in higher dimensional space then apply linearly separable method; this method is costly in computation. (2) The solution of the first method is kernel trick, this method performs dot on original data instead of the transferred data tuple [19-21]. That is

$$K(X_i, X_j = \psi(X_i).(X_j) \qquad (9)$$

Kernel Functions: Kernel functions map the data into better representational space. Some common Kernel functions are

$$K(X_i, X_j) = \begin{cases} X_i.X_J \\ (1 + X_i.X_j)^d \\ \exp(\frac{\| X_i.X_J \|^2}{2\sigma^2}) \\ \tanh(tX_i.X_j + p) \end{cases} \qquad (10)$$

B. K- Nearest Neighbor (K-NN)

K- nearest neighbor classifier predict the class label of an unknown instance by obtaining the K- nearest neighbor's class. The new instance will be the 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), …… (X_n,y_n), where X_i is data points, X_i € R_d, y_i is labeled class corresponding to X_i and let y_i € (+1, -1)
2. Find the D (X_{new}, X_i) 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 X_i.
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 medical datasets are often facing a class imbalance problem. The datasets are called imbalanced if one of the classes contains fewer instances than the other classes [23, 24]. Classifiers mostly produce the biased prediction for the minority classes. Synthetic Minority Oversampling Technique (SMOTE) oversampled the minority class by generating new synthetic observations. By the SMOTE algorithm creations of synthetic observations are based on feature space similarities between existing minorities instances [25-28]can be defined as

D: 'd' dimensional dataset

$S_m \subset D : S_m$ is the set of minority class instance

$x_i \in S_m$: x_i is the minority class instance under consideration

$\delta \in rand(0,1)$

x_{syn} : Synthetic observation or instance

In general, the Synthetic Minority Oversampling Technique contains the following main steps

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

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

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

Step4: Multiply the difference with δ

Step5: Find the new point or observation ' x_{syn} ' on the line segment by adding the obtain value to the feature vector x_i

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

All these steps can be represented simply by the following equation

$$x_{syn} = x_i + (\hat{x}_i - x_i) * \delta \qquad (11)$$

IV. DATASETS

In this study, two datasets of the liver patient have been used to building and testing the models. First dataset has been collected from Madhya Pradesh region of India called Madhya Pradesh Region Liver Patient Dataset (MPRLPD). This dataset consists of 12 important features of Liver Function Test (LFT) which are age, sex, a/g ratio, albumin, alk. phosphatase, direct bilirubin, globulin, indirect bilirubin, sgot, sgpt, total bilirubin, total protein, having two class. This dataset consists of 7865 records in which 6282 persons have the liver disorder and 1583 persons having healthy.

This dataset comprises also information about 5056 males and 2809 females. Description of attributes of MPRLPD is shown in Table 1 Another one is Indian Liver Patient Dataset (ILPD) taken from UCI repository [8] having 11 important features of LFT. This dataset consist 583 patients records in which 416 persons having liver disorder and 167 persons who are healthy.

Table 1. Describes the attributes in MPRLPD

Sl. No	Attribute Name	Type	Range
1	Age	Interval	7-91
2	Gender	Nominal	Male-Female
3	Albumin Globulin Ratio (A/G Ratio)	Interval	0.22-2.05
4	Albumin (ALB)	Interval	1-4.90
5	Alk. Phosphatase (Alkphos)	Interval	38-1592
6	Direct Bilirubin (DB)	Interval	0-40.20
7	Globulin	Interval	2-8
8	Indirect Bilirubin (IB)	Interval	0.10-14.90
9	SGOT	Interval	13-3359
10	SGPT	Interval	16-2231
11	Total Bilirubin (TB)	Interval	0.20-55.10
12	Total Protein (TP)	Interval	3.60-10.20
13	Predictor	Binary	0-1

V. PERFORMANCE MEASURES

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

Confusion Matrix: Confusion matrix summarize the actual and predicted results of a classification model [29]. Confusion matrix identifies the number of correct and incorrect predictions with count values and breaks down into the classes[30] as shown in Table 2.

A. 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.

$$\frac{TP+TN}{TP+TN+FP+FN} \quad (12)$$

B. Specificity (TNR)

The accuracy of the data that classified in negative class.

$$\frac{TN}{TN+FP} \quad (13)$$

C. Sensitivity or Recall (TPR)

Accuracy of the data that classified in Positive class

$$\frac{TP}{TP+FN} \quad (14)$$

D. Precision

It is the number of positive predictions divided by the total number of positive class values.

$$\frac{TP}{TP+FP} \quad (15)$$

E. False Positive Rate (FPR)

Percentage of miss classified (Error) in Negative Class

$$\frac{FP}{FP+TN} \text{ (or) } 1-TNR \quad (16)$$

F. False Negative Rate (FNR)

Percentage of miss classified (Error) in positive class

$$\frac{FN}{FN+TP} \text{ (or) } 1-TPR \quad (17)$$

Table 2. Describes the attributes in MPRLPD

Data Class	Classified as True	Classified as False
True	TP (Correct)	FP(Incorrect)
False	FN (Incorrect)	TN (Correct)

VI. PROPOSED APPROACH

In this paper, we have used Support Vector Machine and K-Nearest Neighbour (KNN) algorithms with or without SMOTE to find the liver disorder form imbalanced Liver Function Test dataset. In this regard, we have used MATLAB R2014a to evaluate the result on MPRLPD and ILPD datasets. The Fig. 1 shows all the steps of implemented work clearly. Implementation of this work follows the following steps-

1. In this regards two datasets have been used. First ILPD dataset has been selected from UCI repository [8] and preprocessed. Second MPRLPD dataset is collected and preprocessed.
2. Perform data balancing of ILPD and MPRLPD imbalance datasets using Synthetic Minority Over-sampling Technique.
3. Train the model using SVM/KNN (on four folds) for both balanced and imbalanced dataset of ILPD/MPLPD.
4. Predict the label on the rest one test set using SVM/KNN train model.
5. Find average results on different parameters of independent test sets and compare the result of an imbalanced and balanced dataset of ILPD/MPLPD.

VII. RESULT AND DISCUSSION

In this study, SVM and KNN(for k=3) algorithms have been applied with oversampling technique. 10-fold cross-validation has been used in order to get the unbiased result. The detail outcomes of the performance matrices are shown in Tables 3 and Table 4. According to Table 3, it can be seen that SVM on MPRLPD give the better result for the parameter accuracy 96.42%, specificity 94.39%, precision 94.12%, and FPR 5.61% with the balanced dataset whereas sensitivity 97.5% and FNR 2.47% with the imbalanced dataset.

Table 3 also shows that the SVM on ILPD, produce the best result for the accuracy 73.96%, specificity 70.96%, precision 65.15%, FPR 29.41% and FNR 3.35% with the balanced dataset whereas sensitivity 86.14% and FNR 13.86% with the imbalanced dataset.

Fig. 3 and Fig. 4 show a comparative graphical representation of performance for imbalance and balance dataset using SVM.

However, according to Table 4, we found that KNN on MPRLPD 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.

Table 3. Performance Measure using SVM

Performance Metrices	SVM on MPRLPD		SVM on ILPD	
	Imbalanced Dataset	Balanced Dataset	Imbalanced Dataset	Balanced Dataset
Accuracy	92.38	96.42	65.21	73.96
Specificity	73.09	94.39	43.63	70.59
Sensitivity	97.53	96.65	86.14	78.98
Precision	90.90	94.12	61.33	65.15
FPR	26.91	5.61	56.37	29.41
FNR	2.47	3.35	13.86	21.02

Table 4. Performance Measure using KNN

Performance Metrices	KNN on MPRPD		KNN on ILPD	
	Imbalanced Dataset	Balanced Dataset	Imbalanced Dataset	Balanced Dataset
Accuracy	76.15	81.42	65.12	74.67
Specificity	38.66	74.72	37.04	70.44
Sensitivity	83.51	93.11	73.79	81.43
Precision	87.43	67.85	79.34	64.15
FPR	61.34	25.28	62.96	29.56
FNR	16.49	6.89	26.21	18.57

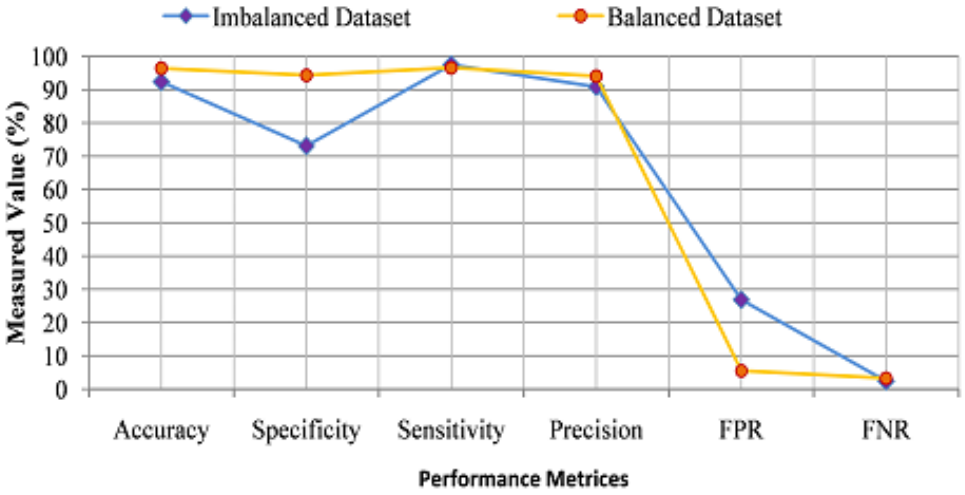

Fig. 2. A comparison between the performance of imbalance and balance dataset of MPRLPD using SVM on different parameters

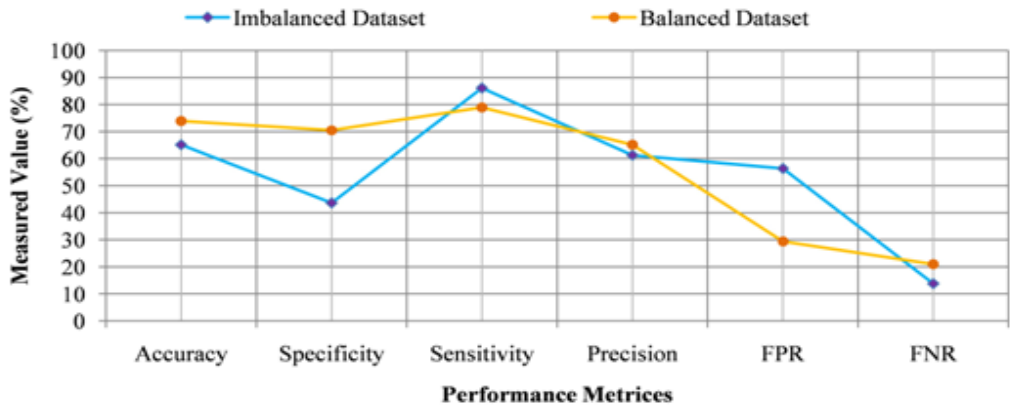

Fig .3. A comparison between the performance of imbalance and balance dataset of ILPD using SVM on different parameters

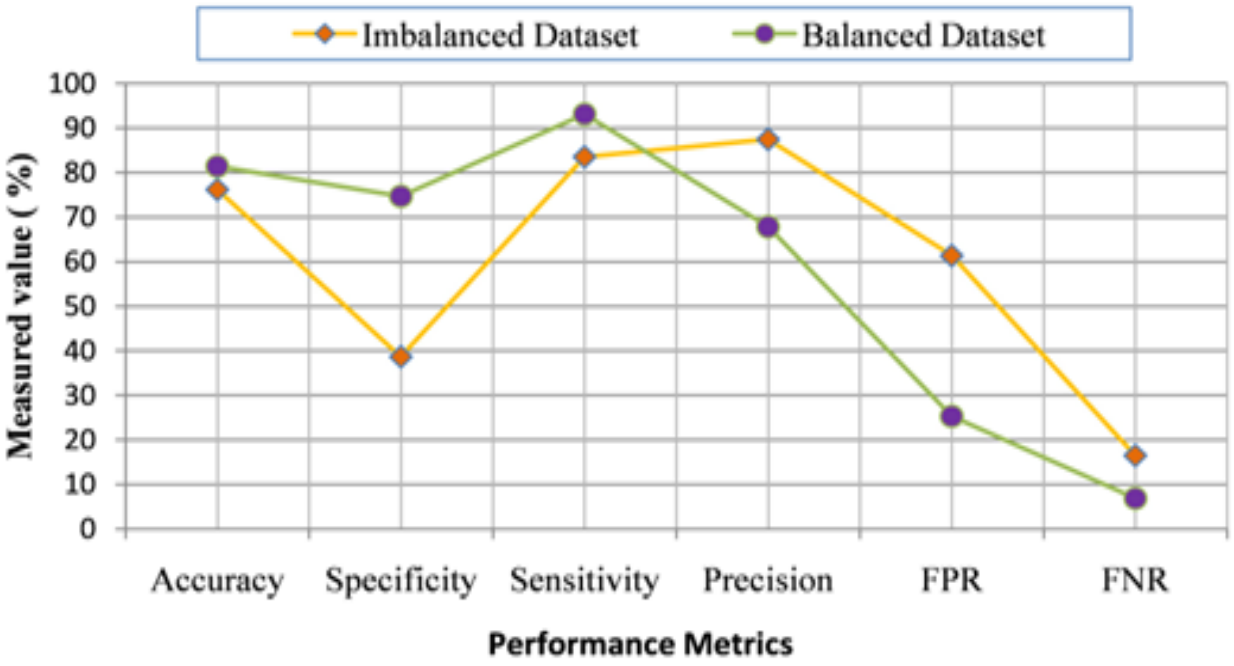

Fig .4. A comparison between the performance of imbalance and balance dataset of MPRLPD using K-NN on different parameters

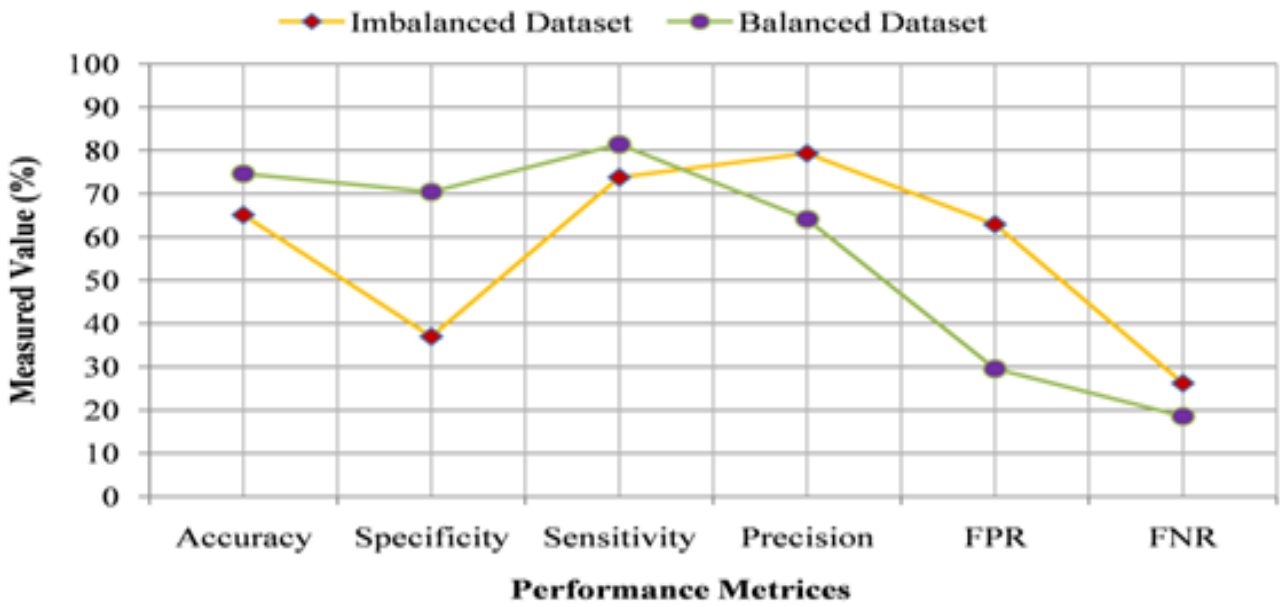

Fig .5. A comparison between the performance of imbalance and balance dataset of ILPD using K-NN on different parameters

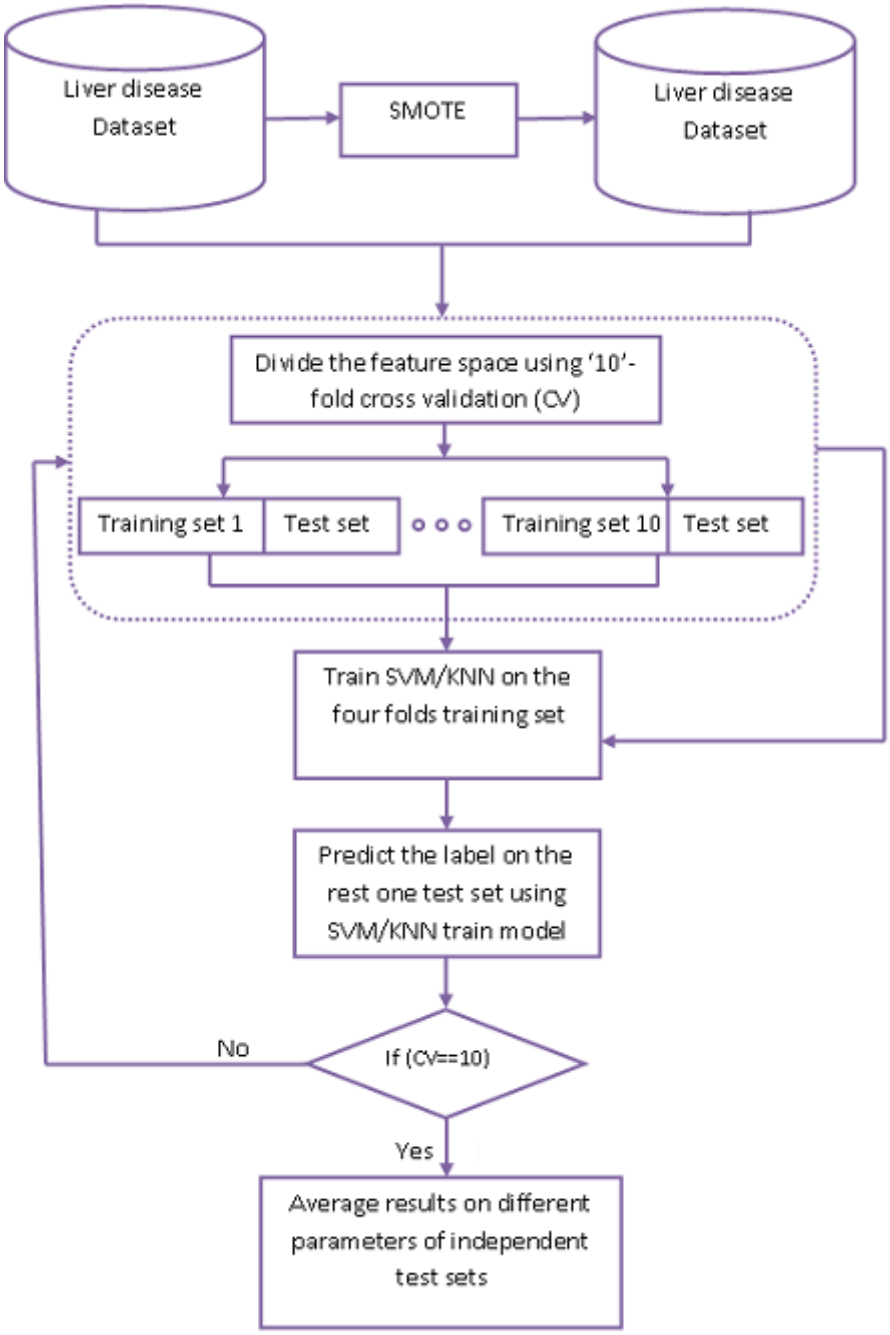

Fig. 1. Flowchart of Proposed Approach

VIII. CONCLUSION

In this paper, we have worked on the imbalanced liver function test dataset to find liver disorder. Imbalanced datasets always provide unfavorable accuracies across the classes of the dataset. So that we have applied a synthetic minority oversampling technique to balance the dataset. In this regard, two well-known algorithms SVM and KNN are applied on both imbalance or balance dataset of ILPD and MPRLPD. Our proposed system shows the improved result on balance dataset with most of the parameter.

REFERENCES

1. J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques. Elsevier, 2011.
2. M. Abdar, M. Zomorodi-Moghadam, R. Das, and I.-H. Ting, "Performance analysis of classification algorithms on early detection of liver disease," Expert Systems with Applications, vol. 67, pp. 239-251, 2017.
3. M. Hassoon, M. S. Kouhi, M. Zomorodi-Moghadam, and M. Abdar, "Rule Optimization of Boosted C5. 0 Classification Using Genetic Algorithm for Liver disease Prediction," in Computer and Applications (ICCA), 2017 International Conference on, 2017, pp. 299-305: IEEE.
4. K. Nagaraj and A. Sridhar, "NeuroSVM: A Graphical User Interface for Identification of Liver Patients," arXiv preprint arXiv:1502.05534, 2015.

5. J. Hopkins. (11/05/2018). Liver: Anatomy and Functions.
6. B. S. A. Benjamin Wedro. (11/05/2018). Liver Disease Facts.
7. K.-C. Cheng, W.-Y. Lin, C.-S. Liu, C.-C. Lin, H.-C. Lai, and S.-W. Lai, "Association of different types of liver disease with demographic and clinical factors," Biomedicine, vol. 6, no. 3, 2016.
8. M. S. P. B. a. N. B. V. Bendi Venkata Ramana. Machine Learning Repository [Online].
9. S. Bahramirad, A. Mustapha, and M. Eshraghi, "Classification of liver disease diagnosis: a comparative study," in Informatics and applications (ICIA), 2013 second international conference on, 2013, pp. 42-46: IEEE.
10. M. ABDAR, "A survey and compare the performance of IBM SPSS modeler and rapid miner software for predicting liver disease by using various data mining algorithms," Cumhuriyet Science Journal, vol. 36, no. 3, pp. 3230-3241, 2015.
11. D. Alemayehu and M. L. Berger, "Big Data: transforming drug development and health policy decision making," Health services and outcomes research methodology, vol. 16, no. 3, pp. 92-102, 2016.
12. S. N. N. Alfisahrin and T. Mantoro, "Data Mining Techniques for Optimization of Liver Disease Classification," in Advanced Computer Science Applications and Technologies (ACSAT), 2013 International Conference on, 2013, pp. 379-384: IEEE.
13. A. Hammad and S. AbouRizk, "Knowledge discovery in data: A case study," Journal of Computer and Communications, vol. 2, no. 05, p. 1, 2014.
14. M. B. Priya, P. L. Juliet, and P. Tamilselvi, "Performance Analysis of Liver Disease Prediction Using Machine Learning Algorithms," 2018.
15. M. Abdar, N. Y. Yen, and J. C.-S. Hung, "Improving the Diagnosis of Liver Disease Using Multilayer Perceptron Neural Network and Boosted Decision Trees," Journal of Medical and Biological Engineering, pp. 1-13, 2017.
16. X. Zhou, Y. Zhang, M. Shi, H. Shi, and Z. Zheng, "Early detection of liver disease using data visualisation and classification method," Biomedical Signal Processing and Control, vol. 11, pp. 27-35, 2014.
17. . V. Ramana, M. P. Babu, and N. Venkateswarlu, "A critical comparative study of liver patients from USA and INDIA: an exploratory analysis," International Journal of Computer Science Issues, vol. 9, no. 2, pp. 506-516, 2012.
18. S. Kant and I. A. Ansari, "An improved K means clustering with Atkinson index to classify liver patient dataset," International Journal of System Assurance Engineering and Management, vol. 7, no. 1, pp. 222-228, 2016.
19. C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273-297, 1995.
20. R. Saxena. (2017, 11/05/2018). SVM CLASSIFIER, INTRODUCTION TO SUPPORT VECTOR MACHINE ALGORITHM.
21. S. Ray. (2015 11/05/2018). Understanding Support Vector Machine algorithm.
22. R. Saxena. (2016, 11/05/2018). KNN CLASSIFIER, INTRODUCTION TO K-NEAREST NEIGHBOR ALGORITHM.
23. H. Patel and G. S. Thakur, "Classification of imbalanced data using a modified fuzzy-neighbor weighted approach," International Journal of Intelligent Engineering and Systems, vol. 10, no. 1, pp. 56-64, 2017.
24. [24] H. Patel and G. S. Thakur, "Improved Fuzzy-Optimally Weighted Nearest Neighbor Strategy to Classify Imbalanced Data," 2017.
25. [25] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of artificial intelligence research, vol. 16, pp. 321-357, 2002.
26. H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Transactions on knowledge and data engineering, vol. 21, no. 9, pp. 1263-1284, 2009.
27. Y. Xu, C. Wu, K. Zheng, X. Niu, and Y. Yang, "Fuzzy–synthetic minority oversampling technique: Oversampling based on fuzzy set theory for Android malware detection in imbalanced datasets," International Journal of Distributed Sensor Networks, vol. 13, no. 4, p. 1550147717703116, 2017.
28. Z. Zheng, Y. Cai, and Y. Li, "Oversampling method for imbalanced classification," Computing and Informatics, vol. 34, no. 5, pp. 1017-1037, 2016.
29. R. Das and A. Sengur, "Evaluation of ensemble methods for diagnosing of valvular heart disease," Expert Systems with Applications, vol. 37, no. 7, pp. 5110-5115, 2010.
30. J. Brownlee. (2016, 11/05/2018). Confusion Matrix in Machine Learning.

AUTHORS PROFILE

Pushpendra Kumar received the Bachelor and Master degree in Computer Applications from B.N.M.U Madhepura (Bihar) in 2011 and R.G.P.V Bhopal (MP) in 2014 respectively. Currently, he is pursuing Ph.D in Computer Applications from Maulana Azad National Institute of Technology, Bhopal (MP). His area of interest includes Data Mining, Machine Learning and Bioinformatics.

Dr. Ramjeevan Singh Thakur is Associate Professor in the Department of Computer Applications at Maulana Azad National Institute of Technology, Bhopal, India. He is a Teacher, Researcher and Consultant in the field of Computer Science and Information Technology. He earned his Master Degree from Samrat Ashok Technology Institute, Vidisha (M.P.) in 1999 and Ph.D. Degree (Computer Science) From Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal (M.P.) in 2008.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication