Detecting Liver Disease Diagnosis by Combining SMOTE, Information Gain Attribute Evaluation and Ranker

Mutiara Auliya Khadija
Department of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Yogyakarta, Indonesia
mutiaraauliya@mail.ugm.ac.id

Noor Akhmad Setiawan
Department of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Yogyakarta, Indonesia
noorwewe@ugm.ac.id

ABSTRACT

Liver Disease is inflammation of liver organ that cause significant damage to the body and most severely it will cause death. Identifying or diagnosing the liver disease in patient need high concern to determine whether the patient really has the disease or not. Health is also influenced with technology. There are data mining technologies that can be used to determine and detect a disease based on the data. With high accuracy will known early identification of liver patient diagnosis and will increase patient survival rate. This research, are combine of SMOTE for preprocessing, Information Gain Attribute Evaluation and Ranker for feature selection. That methods can improve the accuracy of liver disease diagnosis. It compared with four classification using Naïve Bayes, k-NN, Random Forest and SVM. The best accuracy can we obtained using combination of SMOTE, Information Gain Attribute Evaluation and Ranker using Random Forest classification with result 77.06% in accuracy.

Keywords

Liver Disease, Feature Selection, Classification

I. INTRODUCTION

Liver Disease is one of symptom that attacks liver organ in human body. Liver is the important organ that playing a major role in metabolism and serving several vital function for maintain the chemicals like glucose, balancing so many nutrients, fat, vitamin, cholesterol and hormones. Liver divided into 2 part of the left and right hemi liver [1] Because the liver is a very vital organ, if there is inflammation it will cause significant damage to the body and most severely it will cause death. People affected by liver disease usually nausea, vomiting, right upper quadrant abdominal pain, fatigue and weakness.

On the other hand, identifying or diagnosing a disease is very important to do by a doctor. In certain cases, can not be equated with the diagnosis of other diseases. The truth of an illness diagnosis is important because remembering the actions that will be given later. The population is also growing and causing the complexity of the disease suffered by patient also higher. So that a high degree of accuracy is needed to determine whether the patient really has the disease or not [2].

Health is also influenced with technology. There are data mining and machine learning technologies that can be used to determine and detect a disease based on the data it has [3]. The choice of method in data mining will affect the resulting accuracy. If the method is right, so the detection process will has best accuracy. [4] In this research, focused on liver patient diagnosis. Several researches on data mining methods for liver patient disease diagnosis and identification are applied such as using Decision Tree, C4.5, Random Forest [5] , Support Vector Machine, K-Nearest Neighbor [6] dan Bayesian [7] that has many result.

Recently, some research has focused on data mining methods for detecting liver disease diagnosis in increasing accuracy in classification results based on dataset. To get a high accuracy, the dataset requires several stages of pre-processing data and feature selection. High accuracy of liver patient diagnosis will known early identification and will increase patient survival rate.

In this research, we aim to analyze the use of SMOTE for preprocessing, Information Gain Attribute Evaluation and Ranker for feature selection and will be classified by classification methods of Naive Bayes, k-NN, Random Forest and SVM. The evaluation is performed by measuring the accuracy each classification.

The paper is organized as follows Section I for introduction. Section II describes some corresponding research of Liver Patient, Information Gain Attribute Evaluation and Ranker, and classification methods. In Section III we introduce the proposed method. The implementation and experimental results is presented in Section IV. Finally, in Section V we conclude the research.

II. RELATED WORKS

Several studies on liver disease diagnosis have been done. The classification of liver patient dataset using machine learning such as Naïve Bayes, R48, Random Tree and K-star. The accuracy of the Naïve Bayes algorithm for the liver disease dataset is 60.6%, K-star is 67.2%, J48 is 71.2% and the Random Tree algorithm is 74.2%. The highest given with nominal execution time taken is the Random Tree algorithm of 74.2% accuracy. [1]

On the other studies of liver disease prediction, Naïve Bayes, and FT Tree algorithm are compared. The result tells that the accuracy of Naïve Bayes algorithm is better than the other algorithms. The accuracy of this algorithm is found out to be 72.6624%. [7]

For get higher accuracy required feature selection, Pre-processing and classification. Preprocessing using K-means clustering algorithm. Classification using Naive Bayes, AdaBoost, J48, Bagging and Random Forest. A comparative study is performed based on performance measures such as accuracy, error rate, precision, recall and F-measure. Random Forest Algorithm gives best performance 100% accuracy. [8]

Indian Liver Patient Records dataset has 74% accuracy using Logistic Regression on not only training dtaset but also test dataset. The accuracy of Logistic Regression is better than ANN, C 4.5, KNN, SVM or Naïve Bayes Classification [2]

There are hybrid model analysis for improving prediction accuracy of liver patients in 3 phase. First, using classification algorithms. Second, using feature selection CfsSubset Evaluation and Greedy Stepwise. SVM algorithm is considered as the better performance algorithm before applying feature selection. But, Random Forest has better performance after feature selection. Third, making comparation between the results of classification algorithms. The result, Random Forest with the help of feature selection has accuracy of 71.8696% [9]

Attribute selection has important thing in the data mining. In dataset has more number attributes but not all attributes relevant. The research prove that combination of PART classifier with CfsSubsetEval attribute evaluator performs well in terms of precision, recall, f-measure. This CfsSubsetEval method also reduces the mean absolute error of the PART classifier in german_credit dataset using WEKA.

The studies about liver dataset and use hybrid SMOTE technique to handle imbalanced dataset. Comparation between oversampling, undersampling, and hybrid Synthetic Minority Over-sampling Technique (SMOTE) results are applied to SVM for a high classification rate. SMOTE technique has high classification rate compare than oversampling and undersampling.[10]

III. Method

This section describes the dataset and research framework from the process of preprocessing, feature selection, and classification. In this research use dataset ILPD (Indian Liver Patient Dataset) from UCI Machine Learning Repository, consisting of 583 clinical data and 11 attributes. This dataset contains 416 liver patient records and 167 non liver patient records.

Table 1. Attribute in ILPD dataset

No	Attribute	Data Type
1.	Age	Numeric
2.	Gender	Nominal
3.	Total_Bilirubin	Numeric
4.	Direct_Bilirubin	Numeric
5.	Alkaphost (Alkaline Phosphotase)	Numeric
6.	SGPT (Alamine Aminotransferase)	Numeric
7.	SGOT (Aspartate Aminotransferase)	Numeric
8.	Total_Protein	Numeric
9.	Albumin	Numeric
10.	A/G Ratio (Albumin and Globulin Ratio)	Numeric
11.	is_patient	Numeric

The dataset was collected from north east of Andhra Pradesh, India. This dataset contains 441 male patient records and 142 female patient records. The attributes in the ILPD dataset are displayed in the table 1. And the description of its attribute displayed on table 2 [1].

Based on the related works, this research considers the methods of selecting, preprocessing and classifying to improve performance of liver patient diagnosis. The proposed method using of SMOTE, InfoGain Attribute Evaluationand Ranker for preprocessing and will be classified by classification methods of Naive Bayes, k-NN, Random Forest and SVM.

Table 2. Attribute with the Description

No	Attribute	Attribute Description
1.	Age	Age of the patient
2.	Gender	Gender of the patient
3.	Total_Bilirubin	Total Bilirubin
4.	Direct_Bilirubin	Direct Bilirubin
5.	Alkaphost	Alkaline Phosphotase
6.	SGPT	Alamine Aminotransferase
7.	SGOT	Aspartate Aminotransferase
8.	Total_Protein	Total Protein present in patient
9.	Albumin	Albumin amount of the patient
10.	A/G Ratio	Albumin and Globulin Ratio
11.	is_patient	The data belongs to Liver disease patient or not

Overall, the process of this research framework is illustrated in Fig. 1.

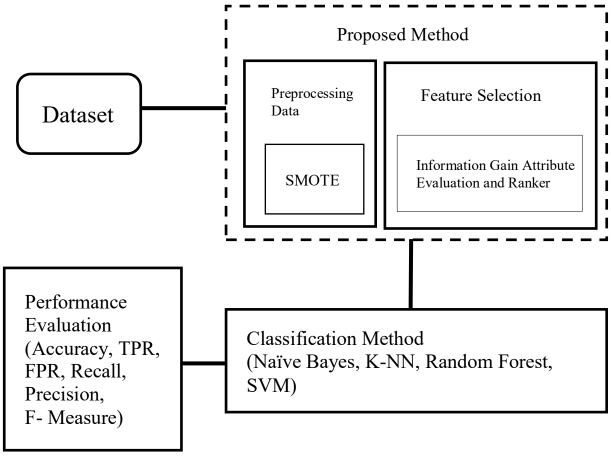

Proposed Method
Dataset
Preprocessing Data
SMOTE
Feature Selection
Information Gain Attribute Evaluation and Ranker
Performance Evaluation (Accuracy, TPR, FPR, Recall, Precision, F- Measure)
Classification Method (Naïve Bayes, K-NN, Random Forest, SVM)

Fig. 1 Proposed Method

A. Preprocessing Data

In this research, preparing the data is the first step of a series of processes to obtain the appropriate data which are used for further management process. In this research used SMOTE for handle the imbalanced dataset. The imbalanced dataset in classification happen when the number of instances that represents one class larger than the other. For this case, can be used sampling technique [11] Sampling can alter imbalanced dataset. There are 2 type of sampling are under sampling and oversampling. Under sampling used for removing instances in set of majority class. And oversampling used for add the instances of minority class. Example of oversampling is SMOTE. [10]

Synthetic Minority Oversampling Technique (SMOTE) increases the minority class by generating new "synthetic" instances based on its number of nearest neighbors [12] The number of the synthetic instances created was set to a number that balances the two classes.[13] SMOTE as "defacto" standard in the framework of learning from imbalanced data due to its simplicity and robustness.[14] In SMOTE the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors [15]

B. Feature selection

Feature selection is one of important parts in pre-processing data before the classification process. It reduces the number of features according to the number of the target class, reduces irrelevant features, excessive features and redundant data that makes the error of the target class. This has a direct effect on the application. The main purpose of choosing a feature is to choose the best feature of a set of data features. In this research using Information Gain Attribute Evaluation and Ranker.

Feature selection also named Attribute Selection, Instance Selection, Data Selection, Feature Construction, Variable Selection or Feature Extraction. Feature selection used for data reduction by redundant and removing irrelevant data to increases the accuracy of data mining algorithms. Feature Selection selects number of relevant features from the original features. [16] Basically, feature selection can be considered as a search problem with some evaluation criteria. One evaluation algorithm is Attribute Evaluation using Information Gain. In WEKA, Information Gain Attribute Evaluation using InfoGainAttributeEval. Search algorithms are necessary for feature selection because it provides a way to search for attributes. For the search algorithm, Information Gain Attribute evaluation should use Ranker. In WEKA, Ranker using weka.attributeSelection.Ranker. [17]

Information Gain Attribute evaluation (IG) is method for measures the significance of attribute by the measure of information gain calculated with respect to target class. This algorithm sets a threshold value and attributes that are above the threshold will be considered for further processing [16] The formula for calculate Information Gain Attribute Evaluation is [18],

$$\text{InfoGain (Class,Attribute)} = \text{H(Class)} - \text{H(Class | Attribute)} \quad (1)$$

C. Classification

Classification is one part of the data mining process, unlike to the cluster algorithm process whose data have no label or target class. So classification can be categorized as supervised learning. In this research, we use four classification algorithms: Naive Bayes, Random Forest, Support Vector Machine and k-Nearest Neighbor [8].

Naive Bayes is a simple probabilistic classification algorithm by computing a set of probabilities based on the sum of frequencies and the value combinations of a dataset. This method only requires a small amount of data in the classification process and often gets unexpected results that do not match the reality.[1]

k-NN is a method of classifying objects based on raster-learning data closest to the object. This method aims to classify new objects based on attributes and training samples. This technique is very simple and easy to implement. Similar to the clustering technique, grouping a new data is based on the new data distance to the nearest data / neighbors.

Random Forest is a classification algorithm that produces the most classes generated by decision trees in which multiple decision trees as classifier and by voting on the available decision trees makes the accuracy increasing,

Support Vector Machine (SVM) is a method that can be used for classification and regression. SVM works best on data with high dimensions. But SVM training time tends to be slow, though SVM is very accurate to handle complex nonlinear models. The weakness of SVM is vulnerable to overfitting when compared to other methods.[19]

D. Performance Evaluation

For the performance evaluation use accuracy, TPR, FPR, Recall, Precision and F-Measure.

Accuracy

It is a ratio of ((no. of correctly classified instances) / (total no. of instances)) *100) and it can be defined as,

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

TPR (True Positive Rate)

True Positive Rate is rate of true positives (instances correctly classified as a given class)

$$TPR = \frac{TP}{TP+FN} \quad (3)$$

FPR (False Positive Rate)

False Positive Rate is simply the ratio of false positives to false positives plus true negatives [20]

$$FPR = \frac{FP}{FP+TN} \quad (4)$$

Recall

Recall is proportion of instances classified as a given class divided by total in that class (equivalent to TP rate)

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

Precision

Precision correctly classified instances belongs to TP divided by number of instances classified as belonging to class. That is, it is the proportion of true positives out of all positive results.

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

F-Measure

F-measure is nothing but combining recall and precision scores into a single measure of performance.

$$F - Measure = \frac{2*recall*precision}{(recall+precision)} \quad (7)$$

IV. RESULT

In this research, use WEKA Tools for doing the proposed method. First, change the data type of *is_patient* attribute from numeric to nominal. In WEKA, use *NumericToNominal* for making nominal type. This attribute consists of liver diagnosis

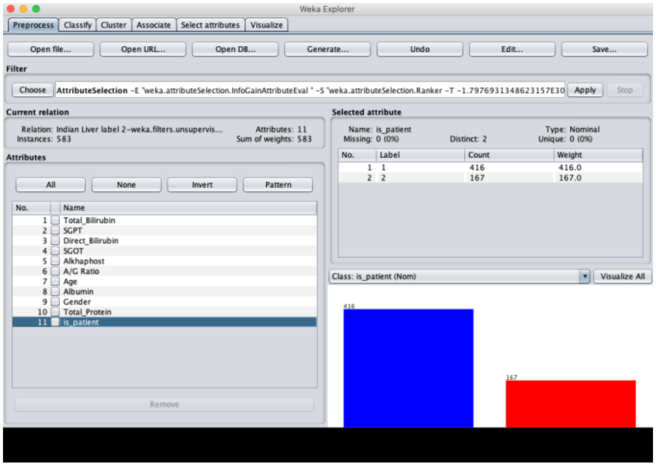
Patient as number 1 and number 2 is for not a patient with liver diagnosis. After that, this experiment conducted four times.

First experiment, the dataset has been classified using Naïve Bayes, k-NN, Random Forest and SVM with cross validation 10. This dataset without preprocessing and feature selection process. It is found that SVM classification was the best performed with an accuracy of 72.38%. The resulted in the performance evaluation of dataset without proposed method presented in Table 3.

Table 3. Results of Classification Without Proposed Method

Method	Accuracy	TPR	FPR	Precision	Recall	F-Measure
Naïve Bayes	55.74 %	0.557	0.206	0.792	0.557	0.560
KNN	64.15 %	0.642	0.466	0.660	0.642	0.649
Random Forest	70.32 %	0.703	0.542	0.673	0.703	0.679
SVM	72.38 %	0.724	0.688	0.801	0.724	0.618

Second experiment, the dataset only used Feature Selection Information Gain Attribute Evaluation and Ranker. In WEKA, this feature selection are InfoGainAttributeEval and weka.attributeSelection.Ranker. So the sequence of the dataset has changed being Total_Billirubin, SGPT, Direct_Billirubin, SGOT, Alkhapost, A/G Ratio, Age, Albumin, Gender, Total_Protein, *is_patient* shows in fig 2.



AttributeSelection -E "weka.attributeSelection.InfoGainAttributeEval " -S "weka.attributeSelection.Ranker -T -1.7976931348623157E30

Relation: Indian Liver label 2-weka.filters.unsupervis... Attributes: 11
Instances: 583 Sum of weights: 583

Name: is_patient Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	1	416	416.0
2	2	167	167.0

No.		Name
1	□	Total_Bilirubin
2	□	SGPT
3	□	Direct_Bilirubin
4	□	SGOT
5	□	Alkhaphost
6	□	A/G Ratio
7	□	Age
8	□	Albumin
9	□	Gender
10	□	Total_Protein
11	□	is_patient

Fig. 2 Attribute After Feature Selection in WEKA

In Table 4, shows the result that SVM still has the best accuracy of 72,38%. The performance increase occurred in the Random Forest classification which originally has the accuracy respectively 70.32% rise to 71.52%.

Table 4. Results of Classification using Feature Selection Information Gain Attribute Evaluation and Ranker

Method	Accuracy	TPR	FPR	Precision	Recall	F-Measure
Naïve Bayes	55.74 %	0.557	0.206	0.792	0.557	0.560
KNN	64.15 %	0.642	0.466	0.660	0.642	0.649
Random Forest	71.52 %	0.715	0.534	0.686	0.715	0.690
SVM	72.38 %	0.724	0.688	0.801	0.724	0.618

Third experiment, dataset are doing the preprocessing data with SMOTE. In SMOTE, the imbalanced dataset will be handled by sampling shows in Table 5. Before using SMOTE, the total data is 583 and after using SMOTE, the total was increase being 750 data.

Table 5. The Total Amount Before and After the SMOTE

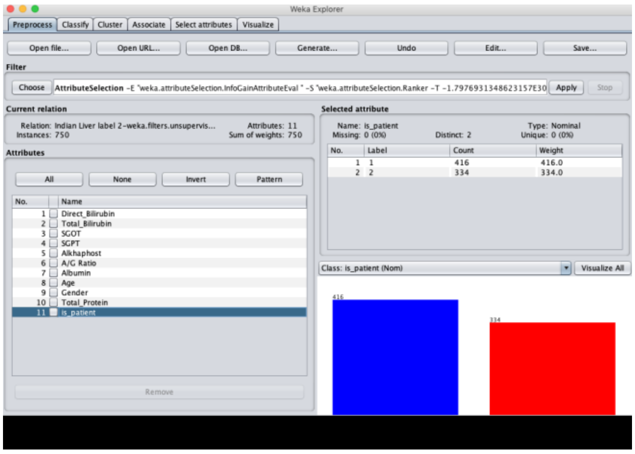
	Liver Patient	Not liver patient	Total
Before SMOTE	416	167	583
After SMOTE	416	334	750

In Table 6, shows results of classification only using Preprocessing SMOTE. The best performance evaluation accuracy in Random Forest classification. The accuracy of Random Forest is 75.46% higher than Naïve Bayes, kNN and SVM. The performance increase occurred in the Naive Bayes classification which originally has the accuracy respectively 55.74% shows on Table 3 rise to 65.06 % shows on Table 6.

Table 6. Results of Classification using Preprocessing SMOTE

Method	Accuracy	TPR	FPR	Precision	Recall	F-Measure
Naïve Bayes	65.06 %	0.651	0.289	0.760	0.651	0.628
KNN	66 %	0.660	0.342	0.663	0.660	0.661
Random Forest	75.46 %	0.755	0.245	0.757	0.755	0.755
SVM	65.88 %	0.659	0.442	0.760	0.659	0.599

Last experiment, using SMOTE for preprocessing and Information Gain Attribute Evaluation and Ranker for feature selection. After that, the dataset classified using Naïve Bayes, k-NN, Random Forest and SVM. From the Feature Selection process, InfoGainAttributeEval measures the significance of attribute by the measure of information gain calculated with respect to target class and use Ranker for search algorithm. So the sequence of the dataset has changed being Direct_Billirubin, Total_Billirubin, SGOT, SGPT, Alkhapost, A/G Ratio, Albumin, Age, Gender, Total_Protein, *is_patient* shows in figure 3.



AttributeSelection -E "weka.attributeSelection.InfoGainAttributeEval " -S "weka.attributeSelection.Ranker -T -1.7976931348623157E30

Relation: Indian Liver label 2-weka.filters.unsupervis... Attributes: 11
Instances: 750 Sum of weights: 750

Name: is_patient Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	1	416	416.0
2	2	334	334.0

No.		Name
1	□	Direct_Bilirubin
2	□	Total_Bilirubin
3	□	SGOT
4	□	SGPT
5	□	Alkhaphost
6	□	A/G Ratio
7	□	Albumin
8	□	Age
9	□	Gender
10	□	Total_Protein
11	□	is_patient

Fig. 3 Attribute After Using Proposed Method in WEKA

In Table 7, shows the result of classification using proposed method. The best accuracy shows in Random Forest classification of 77.06%. And the performance increase occurred in the Random Forest classification which originally has the accuracy respectively 70.32 % shows on Table 3 rise to 77.06 % shows on Table 7

Table 7. Results of Classification using Proposed Method

Method	Accuracy	TPR	FPR	Precision	Recall	F-Measure
Naïve Bayes	65.06 %	0.651	0.289	0.760	0.651	0.628
KNN	66 %	0.660	0.342	0.663	0.660	0.661

Random Forest	77.06 %	0.771	0.231	0.772	0.771	0.771
SVM	65.88 %	0.659	0.442	0.760	0.659	0.599

The confusion matrix of Random Forest shows in Table 8 and graphic ROC (*Receiver Operation Characteristic*) in figure 4.

Table 8. Confusion Matrix of Random Forest

	Liver Patient (Positive)	Not Liver Patient (Negative)
Liver Patient (Positive)	324 (TP)	92 (FP)
Not Liver Patient (Negative)	80 (FN)	254 (TN)

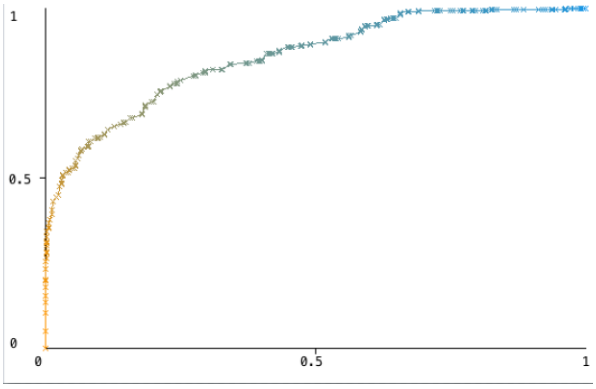


Fig. 4 Graph ROC of Proposed Method with Random Forest

V. Conclusion

In this research, a combination of SMOTE for preprocessing, Information Gain Attribute Evaluation and Ranker for feature selection can improve the accuracy of liver disease diagnosis. It compared with four classification using Naïve Bayes, k-NN, Random Forest and SVM. The best accuracy can we obtained using combination of SMOTE, Information Gain Attribute Evaluation and Ranker also with Random Forest classification with result 77.06% in accuracy. With high accuracy will known early identification of liver patient diagnosis and will increase patient survival rate. For future work, need another combination preprocessing and future selection method for increasing accuracy.

References

[1] S. Muthuselvan, S. Rajapraksh, K. Somasundaram, and K. Karthik, “Classification of Liver Patient Dataset Using Machine Learning Algorithms,” *Int. J. Eng. Technol.*, vol. 7, no. 3.34, p. 323, Sep. 2018.

[2] S. H. Adil, M. Ebrahim, K. Raza, S. S. Azhar Ali, and M. Ahmed Hashmani, “Liver Patient Classification using Logistic Regression,” in *2018 4th International Conference on Computer and Information Sciences (ICCOINS)*, Kuala Lumpur, 2018, pp. 1–5.

[3] R.-H. Lin, “An intelligent model for liver disease diagnosis,” *Artif. Intell. Med.*, vol. 47, no. 1, pp. 53–62, Sep. 2009.

[4] G. R. Krishna, G. V. Ajaresh, I. J. K. Naik, P. R. Dhungyel, and D. K. Prasad, “A New Approach To Maintain Privacy And Accuracy In Classification Data Mining,” vol. 2, no. 1, p. 5.

[5] H. Pakhale and D. K. Xaxa, “A Survey on Diagnosis of Liver Disease Classification,” vol. 2, no. 3, p. 7, 2016.

[6] B. V. Ramana, M. S. P. Babu, and N. B. Venkateswarlu, “A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis,” 2011.

[7] S. Dhamodharan, “Liver Disease Prediction Using Bayesian Classification,” p. 3, 2014.

[8] A. Pathan, “Comparative Study of Different Classification Algorithms on ILPD Dataset to Predict Liver Disorder,” *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 6, no. 2, pp. 388–394, Feb. 2018.

[9] A. Gulia, D. R. Vohra, and P. Rani, “Liver Patient Classification Using Intelligent Techniques,” vol. 5, p. 6, 2014.

[10] K. Lokanayaki and D. Malathi, “Data Preprocessing for Liver Dataset Using SMOTE,” 2013.

[11] M. Hlosta, R. Stríž, J. Kupčík, J. Zendulka, and T. Hruška, “Constrained Classification of Large Imbalanced Data by Logistic Regression and Genetic Algorithm,” *Int. J. Mach. Learn. Comput.*, pp. 214–218, 2013.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[13] S. Jain, E. Kotsampasakou, and G. F. Ecker, “Comparing the performance of meta-classifiers—a case study on selected imbalanced data sets relevant for prediction of liver toxicity,” *J. Comput. Aided Mol. Des.*, vol. 32, no. 5, pp. 583–590, May 2018.

[14] Jie Sun, Hui Li, Hamido Fujita, Binbin Fu, and Wenguo Ai, “Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting,” *Inf. Fusion*, no. 54, pp. 128–144, 2020.

[15] J. Luengo, A. Fernández, S. García, and F. Herrera, “Addressing data complexity for imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary undersampling,” *Soft Comput.*, vol. 15, no. 10, pp. 1909–1936, Oct. 2011.

[16] C. Arun Kumar, M. P. Sooraj, and S. Ramakrishnan, “A Comparative Performance Evaluation of Supervised Feature Selection Algorithms on Microarray Datasets,” *Procedia Comput. Sci.*, vol. 115, pp. 209–217, 2017.

[17] A. O. Balogun, S. Basri, S. J. Abdulkadir, and A. S. Hashim, “Performance Analysis of Feature Selection Methods in Software Defect Prediction: A Search Method Approach,” *Appl. Sci.*, vol. 9, no. 13, p. 2764, Jul. 2019.

[18] Dr Gnanambal S, Dr Thangaraj M, Dr Meenatchi V.T, and Dr Gayathri V, “Classification Algorithms with Attribute Selection: an evaluation study using WEKA,” *Int J Adv. Netw. Appl.*, vol. 09, no. 06, pp. 3640–3644, 2018.

[19] Y. E. Kurniawati, A. E. Permanasari, and S. Fauziati, “Comparative study on data mining classification methods for cervical cancer prediction using pap smear results,” in *2016 1st International Conference on Biomedical Engineering (IBIOMED)*, 2016, pp. 1–5.

[20] P. P. Dhakate, S. Patil, K. Rajeswari, and D. Abin, “Preprocessing and Classification in WEKA Using Different Classifiers,” vol. 4, no. 8, p. 3, 2014.