EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Forecasting Diabetes Correlated Non-alcoholic Fatty Liver Disease by Exploiting Naïve Bayes Tree

Shiva Shankar Reddy[1,*], Nilambar Sethi[2], R. Rajender[3], Gadiraju Mahesh[4]

[1]Research Scholar, Department of C.S.E., Biju Patnaik University of Technology, Rourkela, Odisha, India
[2]Department of Computer Science and Engineering, GIET, Gunupur, Odisha, India
[3]Department of Computer Science and Engineering, LENDI Engineering College, Vizianagaram, India
[4]Department of Computer Science and Engineering, SRKR Engineering College, Bhimavaram, India

Abstract

INTRODUCTION: In recent years, non-alcoholic fatty liver disease (NAFLD) has been identified as the most vulnerable chronic disease. Fat is accumulated in the liver cells of persons with NAFLD. Diabetes is the most common ailment among people of all ages, so it is critical to recognize and prevent its adverse effects.
OBJECTIVES: A relevant dataset with appropriate features was selected. Ensemble algorithms were applied for the prediction task, and finally, the method with the best performance was extracted.
METHODS: In addition to Ensemble approaches namely bagging, Random forest and Ada-boost, individual classifiers Naive Bayes (NB) and C4.5 Decision tree were considered. These ML techniques were compared with the proposed NB tree algorithm, a combination of C4.5 and Naive Bayes.
RESULTS: The following evaluation parameters were computed for each analyzed algorithm: accuracy, detection rate, negative predictive value (NPV), false negative rate (FNR), and false positive rate (FPR). The algorithms are then compared based on these metrics to determine the best algorithm. The NB tree was obtained to be the best method with 97.55% accuracy, 0.4853 detection rate, 0.9615 NPV, 0.0388 FNR, and 0.0099 FPR.
CONCLUSION: The NB tree outperformed individual Naive bayes and C4.5 classifiers, and the other techniques studied. The developed algorithm could be applied in NAFLD-related research.

Keywords: Non-alcoholic fatty liver, diabetes mellitus, ensemble techniques, naive bayes, C4.5 decision tree, bagging, random forest, ada-boost, NB tree, accuracy, detection rate, NPV, FNR and FPR, diabetes mellitus (DM).

Received on 22 January 2022, accepted on 19 April 2022, published on 29 April 2022

doi: 10.4108/eai.29-4-2022.173975

*Corresponding Author. Email: shiva.shankar591@gmail.com

1. Introduction

Diabetes mellitus has recently emerged as one of the most vulnerable chronic diseases. People are becoming more susceptible to this disease due to modern lifestyle factors such as bad food habits and a lack of physical activity. Glucose levels in the blood are not appropriately maintained, leading to various complications. One such problem is a non-alcoholic fatty liver disease (NAFLD). This consequence occurs in diabetic people who consume little or no alcohol. Fat storage is detected in the liver cells of NAFLD patients, and there is a risk of liver injury and inflammation [1].

Usually, the liver contains a limited amount of fat; however, if the amount surpasses the limit, it is called fatty liver. The Figure 1 illustrates the difference in appearance between healthy liver and fatty liver. The liver performs vital activities in the human body. The liver performs various activities, including albumin and clotting factor manufacturing, blood detoxification, nutrient and drug processing, fat, vitamin, bile storage, and glucose

production. As a result, detecting liver illnesses is a crucial responsibility to reduce their negative effects [2].
NAFLD is caused by both type 1 and type 2 diabetes. However, people with type 2 diabetes are more vulnerable to it than people with type 1 diabetes. The alanine amino transferase (ALT) levels of 20% of children with type 2 diabetes are abnormal [3]. According to a study [4], NAFLD affects 50–70% of type-2 diabetic individuals and 50% type-1 diabetic patients. They also discovered that diabetic patients have a higher risk of developing advanced NAFLD than non-diabetic ones.

Figure 1. Visualization of healthy liver and fatty liver

Cirrhosis will develop if NAFLD is not appropriately treated. In this situation, the liver will be harmed. Cirrhosis can result in ascites, hepatic encephalopathy, esophageal vein enlargement, liver malignancy, and end-stage liver failure. End-stage liver failure causes a liver function to decline or halt. Ascites are a condition in which fluid accumulates in the abdomen. Slurred speech, confusion, and tiredness are some of the symptoms of hepatic encephalopathy [1]. As a result, early identification of NAFLD will help avert disease progression.
Some of the symptoms of NAFLD include enlarged blood vessels and spleen, red hands, jaundice, lethargy, and abdominal enlargement. If a person notices these symptoms, he should see a doctor avoid the disease's worsening effects. Obesity, high cholesterol, diabetes, high blood pressure, and any other metabolic syndrome might contribute to NAFLD. Among all diabetic patients, those with the highest risk of developing NAFLD [5] have the highest risk of developing the disease.
For the diagnosis of NAFLD, scans and blood tests for liver function are commonly used. Alkaline phosphatase (ALP), aspartate transaminase (AST), alanine transaminase (ALT), gamma-glutamyl transferase (GGT), Bilirubin, and albumin are the parameters for liver function tests. Liver illness is diagnosed based on Amino transferase, which refers to AST and ALT. They will aid in the detection of hepatocellular damage. Bilirubin is a yellow-colored chemical found in human stool and blood. A high bilirubin level indicates jaundice, a marker of NAFLD or liver disease that includes hepatitis. The liver produces almost 10 grams of albumin per day, and abnormally high albumin levels suggest liver illness. A patient with NAFLD should keep a healthy weight, exercise regularly, and live a healthy lifestyle [6].

Because diabetes is such a common chronic condition, accurate predictions of its side effects, NAFLD, are required. This forecast is helpful for diabetes people who need to start therapy at the appropriate moment. A few ML algorithms are being examined to construct a suitable predictive model in this context. In this paper, an algorithm called NB tree, which is an ensemble of naive Bayes and C4.5 decision tree, is developed. The suggested approach is compared to techniques such as naive Bayes, C4.5 decision trees, bagging, random forest, and adaptive boosting. A comparison study is carried out based on accuracy, detection rate, NPV, FNR, and FPR. For the algorithm's implementation, R programming is used. After persuasion of the findings and comparison analysis, the algorithm with superior predicted performance was obtained. This superior algorithm was advised to produce more accurate and better NAFLD predictions.

2. Literature survey

Reddy et al. [7] predicted the hospital readmission of diabetic patients. A deep belief network, a deep learning technique, was used. In addition, gradient boosting, ada-boost, logistic regression, decision tree, and random forest existing algorithms are also implemented. The proposed deep belief network performed better than the remaining techniques regarding specificity, accuracy, NPV, and precision, with 0.6644, 0.6917, 0.7032, and 0.6814, respectively. But logistic regression performed better only in terms of f1-score (0.7833).
Sarwar et al. [8] implemented machine learning techniques: random forest, naive Bayes, SVM, decision tree, logistic regression, and KNN to predict diabetes. For this purpose, the Pima diabetes dataset was selected, and a percentage split of 70% was applied. The training data extracted from the percentage split is used for implementing techniques. Both KNN and SVM obtained 77% of the highest accuracy compared with other techniques.
Reddy et al. [9] detected diabetes using voting strategy and considered Pima diabetes dataset. Implemented algorithms like decision tree, SMO, naive bayes, adaboost-M1 and SVM on the training data obtained after performing k-fold cross validation technique. Evaluation is done by using the test dataset. After implementing voting strategy on all the algorithms, 95% overall accuracy was observed.
Vijayan and Anjali [10] have implemented naive Bayes, decision tree, SVM, decision stump, and adaboost to predict diabetes. The diabetes dataset from the UCI repository was chosen for developing models. Each algorithm other than adaboost is considered base learners to obtain an individual adaboost model. AdaBoost with decision stump has obtained good accuracy of 80.72% and concluded as the best performing algorithm.
Reddy et al. [11] predicted single or combination of correlated ailments related to diabetes. Retinopathy, cardiovascular, and nephropathy are the ailments of diabetes selected in their work. An RDAD dataset taken from a medical centre was used to predict the disease. The proposed

fuzzy logic along with the k-cross validation technique. Fuzzy logic has obtained 97% overall accuracy with 80 ms computation time and is the best performing technique over other schemes.

Kulkarni et al. [12] considered few machine learning algorithms to predict NAFLD. Decision tree, SVM, logistic regression, random forest, ANN, and gradient boosting algorithms are used. A dataset that has been used is taken from a hospital in Pune. This dataset is related to liver disease patients obtained from electronic health records. Firstly, data cleaning followed by feature selection is implemented, then continues with the algorithm implementation. They have identified that diabetes is the most crucial factor after feature selection. The random forest has better performance with 85% accuracy and 1.0 AUROC.

Reddy et al. [13] reviewed various data mining techniques used for diabetes prediction and correlated ailments. Different research works compared methods like C4.5, image Net, I-SVM, fuzzy, and neuro cognitive. By performing k-cross validation technique, image Net obtained better accuracy. So, this was identified as the best among all other data mining techniques.

Deo and Panigrahi [14] highlighted their work on hepatic steatosis prediction. Hepatic steatosis means fatty liver, which can be caused due to either alcoholic or non-alcoholic consumption. NHANES-III dataset was used in their work with a 70% percentage split. SVM with medium and fine Gaussian, bagging, and boosting techniques like gentle and adaboost are implemented along with the 10-cross validation technique. Gentle boosting tree achieved average accuracy of 79.03%, sensitivity of 75.88%, specificity of 81.86%, and AUC of 0.79 and was recognized as the best.

Chen and Zhao [15] proposed multi-layer random forest (MLRF) to predict fatty liver disease. A real time fatty liver disease dataset was considered in their work. Before implementing the proposed technique, data pre-processing, normalization and dimensionality reduction methods are implemented. This proposed technique is then compared with a few algorithms like SVM, naive bayes, logistic regression, and back propagation NN. MLRF was found as the best algorithm with 98.63% accuracy.

Perveen et al. [16] predicted the risk of NAFLD and progression of disease also by using decision tree. The dataset on which they worked is an electronic medical record data. C4.5 is the decision tree algorithm employed for prediction. It was implemented on both balanced and unbalanced datasets and observed that it performed better for unbalanced dataset with 76.2% accuracy, 66.9% precision, 73.5% recall, 67.6% f-measure, 0.299 MCC, and 73.1% AUROC.

Wu et al. [17] used logistic regression, ANN, naive bayes, and random forest algorithms with 3, 5 and 10-cross validation techniques to predict fatty liver. The real time dataset of fatty liver from a hospital was considered in the work. It contains 577 records in total, where 377 are representing as positive for FLD. The best values of accuracy and AUROC are obtained for random forest with 87.48% and 0.925 respectively when 10-cross validation was performed.

The main aim of Wu et al. [18] work is effectively sense motor imagery using EEG signals for mind and system interface. This work is very useful to patients with motor brain problems. This workillustratestechniqueNB algorithm for analysing brain signals. The results of proposed model arebetter than their counter parts.

Islam et al. [19] applied few ML techniques to develop a predictive model for fatty liver disease (FLD). They performed SVM, ANN, random forest and logistic regression techniques with 10-cross validation technique on a liver patient dataset. It contains 994 records with 533 female and 461 male. Among all the techniques logistic regression has performed better with 76.30% accuracy, 74.10% sensitivity and 64.90% specificity.

Details about objectives, dataset and system architecture for this work is given in section 3. This section is followed by section 4 where all the algorithms used in this work are explained. Among these NB tree proposed algorithm is elaborated and remaining are briefly described. The analysis of obtained results, including its discussion, was provided in section 5. In this section the best performing algorithm was found after proper and valid comparison. The conclusion of this work is provided in section 6, followed by references.

3. Methodology

The details of objectives, dataset, and system architecture of the proposed methodology were described in this section.

3.1. Objectives of the work

Non-alcoholic fatty liver is a disease that most diabetic patients will be affected. So, accurate prediction of this disease is needed, which helps a doctor or physician to make a better decision about patient's condition. Machine learning has been widely used to predict various diseases in recent days. It is also a cost effective method for prediction. Hence, few machine learning algorithms are chosen to predict the disease. The aim of this work is

- To obtain a dataset that helps to develop a best predictive model for non-alcoholic fatty liver disease.
- To use an efficient ensemble algorithm for disease prediction.
- To find out the algorithm with best performance.

The used dataset is described in the following sub section. This paper proposes an ensemble method called NB tree, compared to base algorithms such as naive bayes, C4.5 decision tree, bagging, random forest, and ada-boost techniques. R was used to programme all of these algorithms. Accuracy, detection rate, negative predictive value, false negative rate, and false positive rate are used to analyse and compare algorithms to determine the best performing one. Finally, the NB tree outperformed the others.

3.2. Dataset description

Considered dataset has 18 features and 1022 records. This is a binary classification dataset. The target variable has two classes, NAFLD positive and NAFLD negative. A detailed description of the dataset is given in table 1. The attribute hepatitis B is the liver infection. The attributes ALT, AST, GGT, ALP and albumin are the liver functioning test components. These are enzymes of the liver. The abnormal values of these components indicate fatty liver or any other liver disease. Triglyceride is used to detect cholesterol in a person. The higher levels of TG indicate high cholesterol, which is one of the risk factor for NAFLD. The attributes total Bilirubin, direct Bilirubin and indirect Bilirubin are the terms obtained from bilirubin blood test. The abnormal values in these terms indicate liver disease or liver damage. The histogram plot including the density plot is demonstrated in figure 2. The pink color highlighted bars are representing the histogram for each attribute. The density plot is represented as the dashed line for each of the attribute. It is used for visualizing the distribution of dataset considered in this work.

Table 1. Description of attributes in dataset

Attribute name	Description
Age	Age of the patient in years.
Gender	Gender of the patient, male or female.
Height	Height of the patient in cm.
Weight	Weight of the patient in kg.
BMI	Body mass index (kg/m^2). BMI should lie between 18.5 and 24.9 otherwise the patient is said to be obese.
DM	Whether the patient has diabetic mellitus or not. 0 - no diabetes, 1 – has diabetes.
Hypertension	Whether the patient has hypertension or not. 0 – no hypertension, 1 – has hypertension.
Hepatitis	Whether the patient is tested positive or negative for hepatitis B. 0 – negative, 1 – positive.
ALT	Alanine amino transferase, value is given in IU/L and normal range is between 0 and 45 IU/L
AST	Aspartate amino transferase, value is given in IU/L and normal range is between 0 and 35 IU/L
GGT	Gamma-glutamyl transferase, value is given in IU/L and normal range is between 0 and 30 IU/L
ALP	Alkaline phosphate, value is given in IU/L and normal range is between 30 and 120 IU/L
TG	Triglycerides, based on which cholesterol is detected. Its value is given in mmol/L and normal range is <1.7 mmol/L
TBIL	Total Bilirubin, value is given in µmol/L and normal range is between 1.71 and 20.5 µmol/L
DBIL	Direct Bilirubin, value is given in µmol/L and normal range is < 5.1 µmol/L
IBIL	Indirect Bilirubin, it is calculated as TBIL – DBIL. Its value is given in µmol/L
Albumin	Its value is given in g/L and normal range is between 40 and 60 g/L
NAFLD	Whether the patient has non-alcoholic fatty liver disease or not. 1 – positive, 0 – negative.

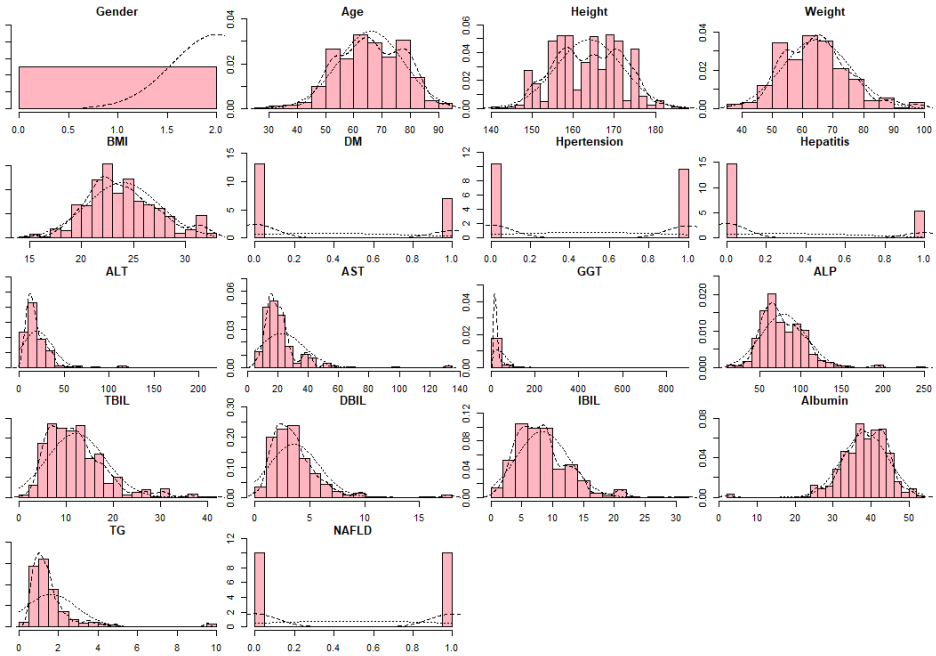

Figure 2. Plot of the dataset

3.3. System architecture

The system architecture from figure 3 demonstrates the working of proposed approach to develop an effective model for NAFLD. Data pre-processing followed by 80% percentage split is performed initially to obtain training and test datasets. On the training dataset with 818 instances all the six algorithms are implemented in R programming. Then a trained model will be obtained for each algorithm further evaluated on the test dataset with 204 instances. This will give the results used to compare all the algorithms in terms of performance metrics accuracy, detection rate, NPV, FNR and FPR. From this analysis, an algorithm with best performance will be found out. In this work, NB tree is recognized as the best performing technique.

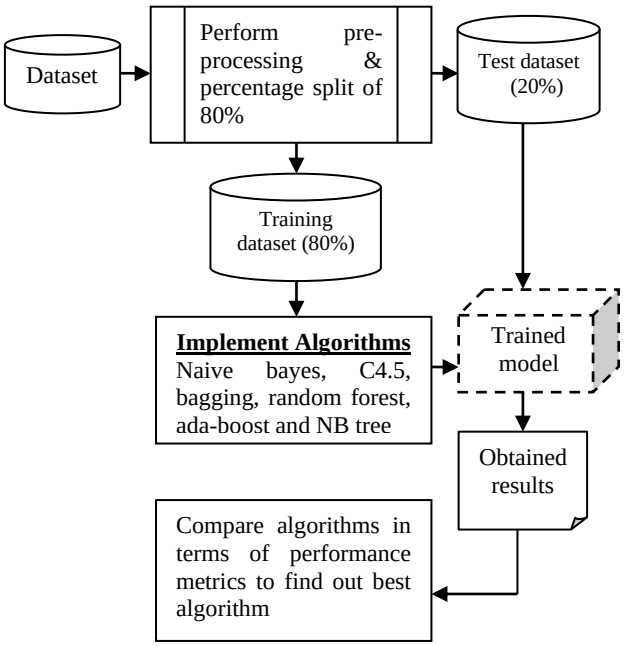

Figure 3. System architecture

4. Algorithms used

NB tree was explained elaborately and remaining five algorithms are briefly described in this section. Remaining algorithms include naive bayes, C4.5 decision tree, bagging, random forest and adaboost.

4.1. Naive bayes classifier

It is a classification algorithm based on probabilistic approach. This algorithm calculates the posterior probability for each group or class of the target variable. This can be briefly explained using formula (1). The term P(G/A) is posterior probability of group G for set of predictor variables A = {a_1, a_2, ...,a_n}. Similarly the terms P(A/G) and P(G) are likelihood and prior probabilities of group G in target variable. The predicted value is the group with highest posterior probability [20] .

$$P\left(\frac{G}{A}\right) = P\left(\frac{A}{G}\right) * P(G);$$
$$\text{here } P\left(\frac{A}{G}\right) = P\left(\frac{a_1}{G}\right) P\left(\frac{a_2}{G}\right) \dots \dots P\left(\frac{a_n}{G}\right) \quad (1)$$

4.2. C 4.5 decision tree

It is also known as J48 decision tree algorithm and basically it is used for classification purpose. The splitting criteria namely gain ratio is employed in this technique for splitting the decision tree until the leaf nodes. Gain ratio is said as the normalization of information gain, which is a splitting criteria employed in ID3 decision tree algorithm. Normalization process is done using split information value. The entire process of constructing a C4.5 decision tree is involved while constructing NB tree. So, this process was explained clearly in NB tree algorithm. The limitation of decision tree is over fitting [21] .

4.3. Bagging

It is an ensemble technique based on weak classifiers. Initially some bootstrap datasets will be constructed. A model trained on a bootstrap dataset is called as a weak learner, whose performance is weak. Such weak learners are combined to get a final model with best performance. Thus, the ensemble technique uses voting strategy to predict the target class. Voting strategy will consider the class which is predicted mostly by weak learners [22] .

4.4. Random forest

It is also an ensemble technique. In this algorithm bootstrap datasets are constructed same as in the bagging technique. Decision tree algorithm was trained on single instance of bootstrap dataset. It also uses voting strategy on the outputs predicted by all the decision trees and gives the predicted value. The value which is predicted mostly after considering all decision trees will be the target value predicted [23] .

4.5. AdaBoost

It's an ensemble technique which works similarly as bagging. In ada-boost the decision stumps are the weak classifiers. Decision stump is a single level decision tree. The difference between bagging and ada-boost is, in ada-boost it assigns weights to the instances for training the decision stumps. Weight of instance will be increased if decision stump training on it is predicted wrongly. The changes in weights are considered for constructing next decision stump. In ada-boost instead of voting the weighted average strategy was used, which performs the average of the weak classifiers. Based on this the final prediction was done by the strong classifier [24] .

4.6. NB tree

The NB tree method combines naive bayes and the C4.5 decision tree. The proposed technique uses a C4.5 or J48 decision tree to build the decision tree. It's an ID3 decision tree that's been tweaked. In the C4.5 decision tree, the attribute selection method for dividing the decision tree is gain ratio. The typical naive bayes algorithm is used at the DT leaves once the DT has been constructed. Each class's probability information for a specific instance is stored in these leaf nodes. The algorithm's result will be the class with the highest probability. This procedure is elaborated in the algorithm given below [25].

Algorithm: NB tree

INPUT: Dataset
OUTPUT: Predictions made for the input data
ASSUMPTIONS: "g" holds the different categories of the target attribute, h holds only one category from g at a time, P_h is the probability of instance that belongs to class h,A is a particular predictor attribute, D is a set of instances from an attribute, k represents different categories of instances in attribute A, p holds a category from k at a time, $|D_p|$ is the no. of instances with category p from attribute A.

Step 1: Start
Step 2: For each attribute A in the input dataset.

a. Calculate entropy for target and predictor attributes.

$$E(D) = -\sum_{h=1}^{g} P_h \log_2(P_h) \quad (2)$$

$$E(D,A) = \sum_{p=1}^{k} \frac{|D_p|}{|D|} E(D_p) \quad (3)$$

b. Calculate Information gain for A.

$$Gain(A) = E(D) - E(D,A) \quad (4)$$

c. Calculate split information value for A.

$$SplitInfo(D,A) = -\sum_{p=1}^{k} \frac{|D_p|}{|D|} \log_2\left(\frac{|D_p|}{|D|}\right) \quad (5)$$

d. Calculate Gain ratio for A.

$$Gainratio(A) = \frac{Gain(A)}{SplitInfo(D,A)} \quad (6)$$

Step 3: Repeat step-2 until all the predictor attributes are completed then end for loop and go to step-4.
Step 4: Predictor attribute which obtained highest value of gain ratio is selected for splitting criteria.
Step 5: After completing the construction of decision tree perform naive bayes on the leaves of the tree.
Step 6: The target class with highest probability from naive bayes is the predicted output for the given input.
Step 7: Stop

Steps 2 and 3 together represents a FOR loop. Its main motive is to calculate gain ratio for each predictor attribute. The gain ratio is the normalization of information gain, which is a splitting criteria used in ID3. Step 2a calculates entropy for target and predictor attributes using formula (2) and (3) respectively. Step 2b calculates information gain for each predictor attribute using formula (4). Steps 2c and 2d calculates split information using formula (5) and gain ratio using formula (6) respectively. Step 4 comprises of identifying attribute with highest gain ratio for splitting the tree. In step 5 naive bayes algorithm is implemented on the decision tree leaves. The formula used to calculate probability for each class is provided in formula (1). The target class with highest probability will be the predicted output from NB tree [26].

5. Results analysis & Discussion

This section contains the outcomes of implementing the discussed strategies in R programming. The computed results are discussed, and all algorithms are compared to determine the best algorithm. The confusion matrix for the NB tree is shown in Table 2. The true positive, true negative, false positive, and false negative values for the NB tree are 99, 100, 1, and 4 correspondingly, as shown in table 2. These values are used to compute the evaluation parameters, which are shown in the subsection that follows. The remaining algorithms are similarly evaluated using a similar way. After that, a comparison is done to determine which algorithm is better.

Table 2. Confusion matrix of NB tree

		Predicted values	
		Positive	Negative
Actual values	Positive	99	1
	Negative	4	100

5.1. Performance metrics

Accuracy

This metric will measure the correct classification rate. The ratio of correct predictions made to the instances in total is the accuracy, with value between 0 and 1. The value nearer to 1 indicates good performance of the model. The formula (7) is to calculate accuracy of the prediction model and its value is obtained as 0.9755 for NB tree. This is 97.55% which can be said as a good performance.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

Accuracy of NB tree = (99+100) / (99+100+1+4) = 0.9755

Detection rate

This metric will measure the ability of a model, to detect different groups in the target variable. The ratio of correct

positive predictions made to the instances in total is called detection rate. Its value is between 0 and 1, where the value near to 1 represents a good performance.

$$DR = \frac{TP}{TP + TN + FP + FN} \quad (8)$$

DR for NB tree = 99 / (99+100+1+4) = 99 / 204 = 0.4853

NPV

The ratio of correct negative predictions made to the actual negative instances in total is called NPV, whose value lies between 0 and 1. The value close to 1 will represent a good performance.

$$NPV = \frac{TN}{TN + FN} \quad (9)$$

NPV for NB tree = 100 / (100+4) = 0.9615

FNR or miss rate

FNR is also called as miss rate. It is the ratio of incorrect predictions made as positive to the total positive predictions, whose value is between 0 and 1. The less value i.e. nearer to 0 indicates good performance.

$$FNR = \frac{FN}{FN + TP} \quad (10)$$

FNR for NB tree = 4 / (4+99) = 0.0388

FPR or fall out

FPR is also called as fall out. It is the ratio of incorrect predictions made as negative to the total negative predictions. Its value lies in between [0, 1]. The value nearer to 0 means good performance.

$$FPR = \frac{FP}{FP + TN} \quad (11)$$

FPR for NB tree = 1 / (1+100) = 0.0099

5.2. Results obtained

Table 3 comprises of obtained results for all algorithms. The proposed ensemble approach NB tree has higher values based on the five metrics analyzed. After comparing naïve bayes and C4.5 decision trees to the NB tree, it was determined that an ensemble of two base classifiers performed better. In addition, three ensemble methods were compared to the NB tree, including bagging, random forest, and adaptive boosting. Though the three ensemble approaches outperformed the individual NB classifiers, the suggested NB tree outperformed them all. Accuracy, detection rate, NPV, FNR, and FPR values for the NB tree are 97.55%, 0.4853, 0.9615, 0.0388, and 0.0099, respectively. The comparison of all the techniques based on each metric is illustrated in figures 4, 5. From the figures it was clear that the proposed NB tree has outperformed individual naive bayes, C4.5 decision tree and other ensemble techniques based on all the five metrics.

Table 4 shows a performance comparison of the algorithms studied in this study with algorithms from relevant literature. Different datasets were utilised in diverse academic publications from table, and the dataset used in this work was different as well. In [12], C4.5 and random forest algorithms are utilised, with random forest proving to be the most effective. In [14], bagging and adaboost were utilised and compared to their rivals, but neither one of them was found to have the best performance. The proposed study took into account the NB classifier, which authors used in works [15] and [17]. Random forest was discovered to be the best algorithm in literary works [12] and [17], and it was also noticed to be applied in [19]. The algorithms employed in the proposed work are drawn from these studies and compared to the proposed NB Tree technique.

Table 3. Results of all algorithms

Algorithm name	Accuracy (%)	DR	NPV	FNR	FPR
Naive Bayes	79.9	0.4412	0.8488	0.1262	0.2772
C4.5	94.61	0.4559	0.9091	0.0971	0.0099
Bagging	93.63	0.4461	0.8929	0.1165	0.0099
Random forest	97.06	0.4804	0.9524	0.0485	0.0099
Ada-Boost	83.82	0.3971	0.8036	0.2136	0.1089
NB tree	97.55	0.4853	0.9615	0.0388	0.0099

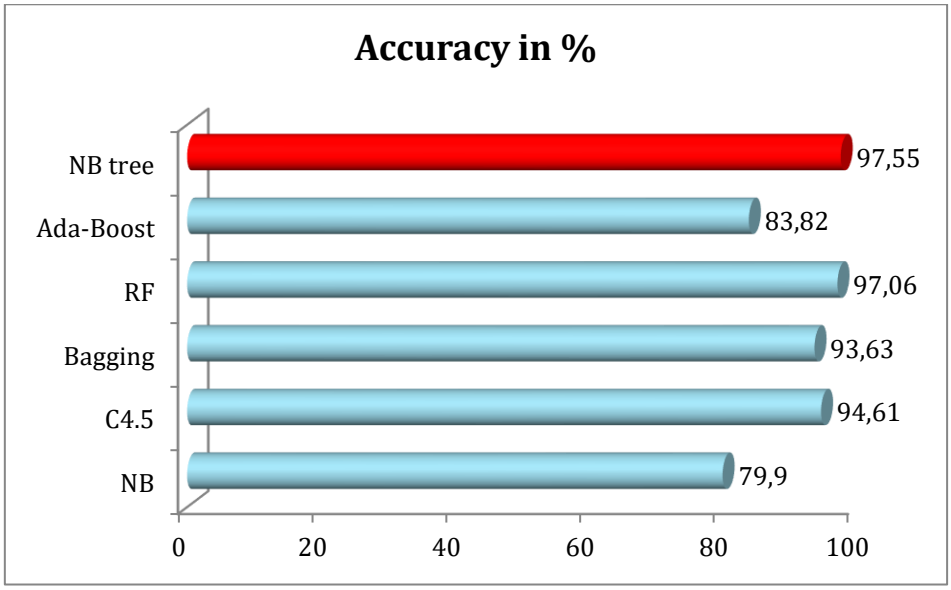

Figure 4. Comparing algorithms in terms of accuracy

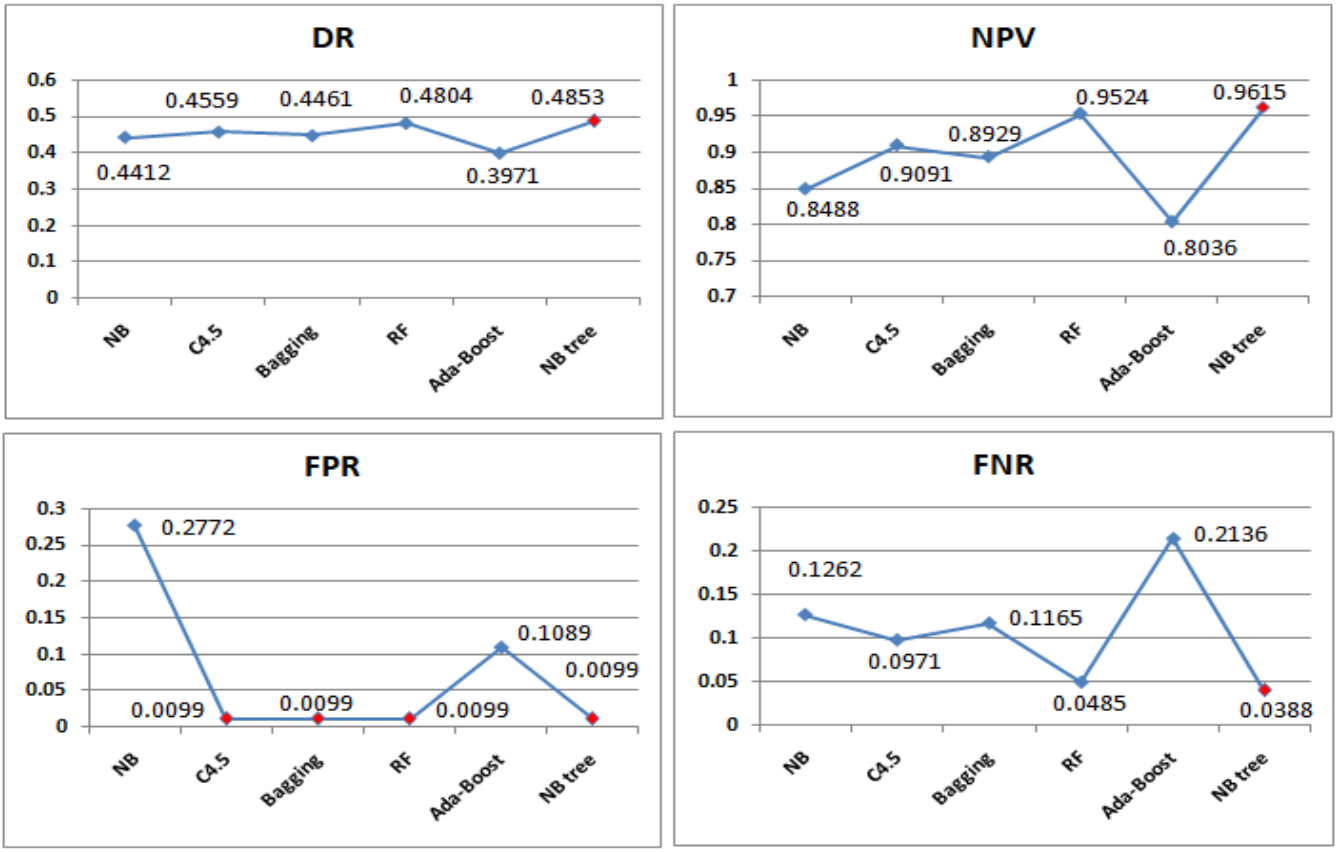

Figure 5. Comparing algorithms in terms of DR, NPV,FPR and FNR

Table 4.Proposed work Vs related literature work

Authors of the work	Used algorithms	Findings in the work	Best performed algorithm	Results (values of metrics)
This work	Naive bayes, C4.5 decision tree, bagging, random forest, adaboost and NB tree.	Best algorithm for prediction of NAFLD is found out by comparing few ensemble techniques. Proposed NB tree algorithm performed most efficiently than remaining in terms of accuracy, NPV, DR, FNR and FPR.	NB tree	97.55% accuracy, 0.9615 NPV, 0.0388 FNR, 0.0099 FPR and 0.4853 DR
Kulkarni et al. [12]	Decision tree, SVM, logistic regression, random forest, ANN and gradient boosting.	Predicted NAFLD using ML techniques. Data cleaning followed by feature selection is performed. Accuracy and AUROC of all algorithms are compared to get best one.	Random forest	85% accuracy and 1.0 AUROC
Deo and Panigrahi [14]	SVM with medium and fine Gaussian, bagging and boosting techniques like gentle and adaboost.	Predicted fatty liver disease by applying 70% percentage split on the dataset. In addition 10-cross validation was also performed after percentage split and compared results in terms of accuracy, sensitivity and specificity.	Gentle boosting tree	Average accuracy-79.03%, sensitivity-75.88%, specificity-81.86% and AUC-0.79
Chen and Zhao [15]	SVM, naive bayes, logistic regression, back propagation neural network and multi-layer random forest (MLRF).	Data pre-processing, normalization and dimensionality reduction techniques are performed sequentially. Then the proposed algorithm MLRF is used to predict fatty liver disease. Evaluation and comparison of results is done using accuracy.	MLRF	Accuracy-98.63%
Perveen et al. [16]	C4.5 decision tree.	Predicting the risk of affecting to fatty liver and its progression is done. Implemented C4.5 on both balanced and unbalanced datasets. Compared the results in those two cases using accuracy, precision, recall, f-measure, MCC and AUROC.	C4.5 decision tree with unbalanced dataset	Accuracy-76.2%, precision-66.9%, recall-73.5%, f-measure-67.6%, MCC-0.299 and AUROC-73.1%.
Wu et al. [17]	Logistic regression, ANN, naive bayes and random forest.	Implemented four ML techniques to predict FLD. All these techniques are implemented using 3, 5 and 10 cross validation. The results of algorithms in case of three cross validation techniques are compared on the basis of accuracy and AUROC.	Random forest	Accuracy-87.48% and AUROC-0.925
Islam et al. [19]	SVM, ANN, random forest and logistic regression.	Developed a predictive model for fatty liver using 10 cross validation with algorithms. For evaluating and comparing the algorithms accuracy, sensitivity and specificity are used.	Logistic regression	Accuracy-76.30%, sensitivity-74.10% and specificity-64.90%

As any of the work from the literature has not used NB tree, the comparison of existing algorithms with it will help to recognise the most significant algorithm for predicting NAFLD. From the result analysis, it had been found that the proposed NB tree outperformed its individual classifiers and other ensemble techniques from literature works. Hence, it was undoubtedly the best algorithm which has obtained about 97.55% accuracy. This value of accuracy is better than accuracy of existing algorithms in the related literature works as well. Also the dataset features in this proposed work also plays a vital role for prediction as they are related to test results from liver function test and few risk factors of NAFLD. Henceforth, it would be better to consider NB tree over other three ensemble algorithms in further works related to NAFLD.

6. Conclusion

Diabetes type 1 and 2 people are more likely to suffer from non-alcoholic fatty liver disease than those who do not have the disease. Clinicians can use the desease prediction to minimize more difficulties to make quick and efficient treatment decisions. The NB tree, an ensemble of naive bayes and a C4.5 decision tree, is the best method for this task. This model was superior to random forest, bagging, and adaboost in terms of accuracy. Naive Bayes and C4.5 are also compared with the NB tree that is an ensemble of these underlying algorithms. After a rigorous comparitive study, NB tree is identified as better performing algorithm with accuracy, detection rate, NPV, FNR and FPR of 97.55 percent , 0.4853, 0.9615, 0.0388 and 0.0099 accordingly. Finally, the NB tree was found superior to other ensembles in predicting NAFLD following a valid and fair review of all outcomes. It is expected to see a lot more work in the medical field in the future employing the best mix of mining algorithms and ML algorithms.

References

[1] Non alcoholic fatty liver disease [online]. Mayo Clinic; [cited 2020 October 21]. Available from: https:// www.mayoclinic.org/diseases-conditions/nonalcoholic-fatty-liverdisease/symptoms-causes/syc-20354567

[2] Davis CP, Shiel WC. Liver Blood Tests [online]. Medicine Net; [cited 2020 June 22]. Available from: https://www.medicinenet.com/liver_blood_tests/article.htm #what_are_the_basic_functions_of_the_liver

[3] Bhatt HB, Smith RJ. Fatty liver disease in diabetes mellitus. Hepatobiliary Surg Nutr. 2015; 4(2):101-108.

[4] Singh A. Risk of non-alcoholic fatty liver disease in patients with type-1 diabetes [online]. ATLAS of Science; [cited 2019 Feb 18]. Available from: https://atlasofscience.org/risk-of-non-alcoholic-fatty-liver-disease-in-patients-with-type-1-diabetes/

[5] Symptoms & Causes of NAFLD & NASH What are the symptoms of NAFLD and NASH? [online]. NIH: National Institute of Diabetes and Digestive and Kidney Diseases; [cited 2016 November]. Available from: https://www.niddk.nih.gov/health-information/liver-disease/nafld-nash/symptoms-causes

[6] Lala V, Goyal A, Bansal P, Minter, DA. Liver Function Tests [online]. StatPearls [online]; [cited 2020 July 4]. Available from: https://www.ncbi.nlm.nih.gov/books

[7] Reddy SS, Sethi N, Rajender R. Evaluation of Deep Belief Network to Predict Hospital Readmission of Diabetic Patients. In: Proceedings of 2020 Second International Conference on Inventive Research in Computing Applications ; 2020; Coimbatore, India. IEEE; 2020. p. 5-9.

[8] Sarwar MA, Kamal N, Hamid W, Shah MA. Prediction of Diabetes Using Machine Learning Algorithms in Healthcare. In: Proceedings of 24th International Conference on Automation and Computing (ICAC); 2018; Newcastle upon Tyne,United Kingdom. IEEE; 2018. p. 1-6.

[9] Reddy SS, Rajender R, Sethi N. A data mining scheme for detection and classification of diabetes mellitus using voting expert strategy. International J. of Knowledge-Based and Intelligent Engineering Systems. 2019; 23(2):103-8.

[10] Vijayan VV, Anjali C. Prediction and diagnosis of diabetes mellitus - A machine learning approach. In: Proceedings of Recent Advances in Intelligent Computational Systems ; 2015; Trivandrum, India. IEEE; 2015. P. 122-127.

[11] Reddy SS, Sethi N, Rajender R. Mining of multiple ailments correlated to diabetes mellitus. Evolutionary Intelligence. 2020; 1-8.

[12] Kulkarni A, Shinde S, Kadam D. Automated Prediction of Non Alcoholic Fatty Liver Disease using Machine Learning Algorithms. International Research J. of Engineering and Technology (IRJET). 2020; 7(9):488-491.

[13] Reddy SS, Sethi N, Rajender R. A Review of Data Mining Schemes for Prediction of Diabetes Mellitus and Correlated Ailments. In: Proceedings of 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA); 2019; Pune, India. IEEE; 2019. p. 1-5.

[14] Deo R, Panigrahi S. Prediction of Hepatic Steatosis (Fatty Liver) using Machine Learning. In: Proceedings of the 2019 3rd International Conference on Computational Biology and Bioinformatics; 2019; ACM. p. 8-12.

[15] Chen M, Zhao X. (2018). Fatty Liver Disease Prediction Based on Multi-Layer Random Forest Model. In: Proceedings of the 2nd Int. Conference on Computer Science and Artificial Intelligence; 2018; ACM. p. 364-368.

[16] Perveen S, Shahbaz M, Keshavjee K, Guergachi A. A Systematic Machine Learning Based Approach for the Diagnosis of Non-Alcoholic Fatty Liver Disease Risk and Progression. Scientific Reports. 2018; 8(2112).

[17] Wu CC, Yeh WC, Hsu WD, Islam M, Nguyen PA, Poly TN, Wang YC, Yang HC, Li YC. Prediction of fatty liver disease using machine learning algorithms. Computer Methods and Programs in Biomedicine. 2019; 170:23-29.

[18] Wang H, Zhang Y. Detection of motor imagery EEG signals employing Naïve Bayes based learning process. Measurement. 2016 May 1;86:148-58.

[19] Islam MM, Wu CC, Poly TN, Yang HC, Li YJ. Applications of Machine Learning in Fatty Live Disease Prediction. Studies in Health Technology and Informatics. 2018; 247:166-170.

[20] Kapoor S, Verma R, Panda SN. Detecting Kidney Disease using Naive Bayes and Decision Tree in Machine Learning. International J. of Innovative Technology and Exploring Engineering (IJITEE). 2019; 9(1):498-501.

[21] Bashir S, Qamar U, Khan FH, Javed MY. An Efficient Rule-based Classification of Diabetes Using ID3, C4.5 & CART Ensembles. In: Proceedings of 12th International Conf. on Frontiers of Information Technology; FIT; 2014.

[22] Perveen S, Shahbaz M, Guergachi A, Keshavjee K. Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. In: Proceedings of Procedia Computer Science; 2016; 82:115 – 121

[23] VijiyaKumar K, Lavanya B, Nirmala I, Caroline SS. Random Forest Algorithm for the Prediction of Diabetes. In: Proceedings of 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN); 2019; Pondicherry, India. IEEE; 2019. p. 1-5.

[24] Chen P, Pan C. Diabetes classification model based on boosting algorithms. BMC bioinformatics. 2018; 19:109.

[25] Devi TS, Sundaram KM. A Comparative Analysis of Meta and Tree Classification Algorithms using Weka. International Research J. of Engineering and Technology (IRJET). 2016; 3(11):77-83.

[26] Mahmood DY, Hussein MA. Analyzing NB, DT and NB Tree Intrusion Detection Algorithms. Journal of Zankoy Sulaimani- Part A (JZS-A). 2014; 16(1):87-94.