IJNRD.ORG ISSN : 2456-4184

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

International Research Journal IJNRD Research Through Innovation

LIVER DISEASE PREDICTION USING DECISION STUMP AND NAÏVE BAYE'S ALGORITHMS (ENSEMBLING TECHNIQUE)

P.Spandana Valli[1], S.Padma Sri[2], E.VijayaDurga[3], T.Suneel Kumar[4], D.J Abhishek[5]

[1]Assistant Professor, Department of Computer Science and Engineering, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, India.

[2,3,4,5] Students , Department of Computer Science and Engineering, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, India.

ABSTRACT

One of the most serious illnesses that can affect the human body is liver cancer. To save human lives, this illness must be diagnosed as soon as possible. Eliminating waste produced by creatures and preserving essential chemicals needed by the body are the liver's two main functions. Around the globe, liver disease affects 4.5% to 9.5% of individuals. Currently, liver function blood test and scan results are analysed to identify illnesses linked to the liver. It is more costly and takes longer time. Machines can execute a job repeatedly without getting tired, but humans do. As a result of technological advancements, machine learning algorithms can now anticipate the onset of illness. The Indian Liver Patient dataset, which has 583 cases and 11 attributes, was obtained from the Kaggle database. The information is randomly split into 507 and 76 intsances, which are then used to train and test the model, respectively. On a sample of 76 occurrences, the suggested machine learning model produced accuracy of 94.7%.

Keywords: Machine Learning, Data mining, Decision Stump, Liver Disease, Ensembling Technique

1. INTRODUCTION

The liver is a large tissue the size of a football. It is important for survival because it conducts many critical biological functions such as protein production and blood clotting, cholesterol, glucose (sugar) and iron metabolism, bile production and excretion, eliminating toxins from the body, blood detoxification and purification. The absence of these processes can cause serious harm to the body. It could be inherited. It can be triggered by a number of variables, including virus infection, obesity, and alcohol intake. Hepatotrophic viruses such as hepatitis A,B,C,D, and E can induce liver illness. There are two types of liver disease: acute and persistent. Acute liver failure can strike suddenly in a matter of days or weeks, typically in the absence of any pre-existing conditions. As previously stated, it is caused by the hepatitis virus or medications. Chronic liver dysfunction is more prevalent. This acute liver disease lasts less than 26 weeks (six months), whereas chronic liver disease causes gradual decline for more than six months.

Liver illness does not always manifest as visible indications and symptoms. The following signs of liver disease are present: yellowish skin and eyes (jaundice), stomach pain, puffiness in the legs and feet, easy bruising, nausea, vomiting, loss of hunger, itchy skin, dark urine color, pallid stool,weight loss,fatigue, backpain,enlarged spleen and gall bladder.

The project's goal is to determine whether an individual has liver disease based on various parameters such as bilirubin, direct bilirubin, total proteins, sgpt, sgot, albumin, and globulin ratio numbers. Currently, liver illnesses are found by analysing liver function blood test results and scan reports, which requires more time and money. From the patient's perspective, it is extremely difficult because he or she must attend numerous check-up sessions and undergo numerous tests in order to receive proper and precise findings. This research employs various machine learning algorithms, such as Naïve Bayes and Decision stump, to provide a decision support model that can assist the physician in predicting liver illness from the dataset. The following is how the document is structured: The related works are described in Section 2. Section 3 describes the methodology used in this study. Section 4 presents the system architecture. Testing and Results are described in Section 5 . Section 6 concludes the summary of the report.

2. LITERATURE SURVEY

Machine learning is a tool used in health care to assist medical workers in patient treatment and clinical data handling. It has attracted a large number of researchers and has been used in a variety of disciplines around the globe. Machine learning has proven its strength in medicine, where it has been used to handle many critical situations such as cancer treatment, heart disease diagnosis, and so on. Many studies have used Decision Tree algorithms, which are one of several excellent techniques. Liver disease in humans has been steadily growing as a result of harmful gas inhalation, contaminated food consumption, various types of medications, and excessive alcohol consumption. Doctor’s workload could be reduced if automatic categorization is used. To cleanse, the pre-processing technique is used. After cleansing the data, a pre-processing technique is used to clean it up for effective categorization. The dataset includes 15 traits associated with actual medical data. Dr.A.S. AneeshKumar oversaw the project Estimating the surveillance of liver disorder using classification algorithms in 2012, which employed a technique for successful categorization of liver disease datasets. His research employed two algorithms: C4.5 and Naive Bayes. According to the findings of his research, C4.5 provides greater precision and Naive Bayes. However, Naive Bayes is sometimes superior to the FP growth method. Later, many algorithms are used to screen for liver illness.[1]

Similarly, Nazmun Nahar and Ferdous Ara released a journal in 2018 on Liver disease prognosis using various decision tree methods. They arrive at the opinion that the techniques were assessed and their success was compared. Based on that study, Decision Stump beats other algorithms, with an accuracy of 70.67%. As a result, using a decision tree to forecast liver disease will help people manage their health.[2]

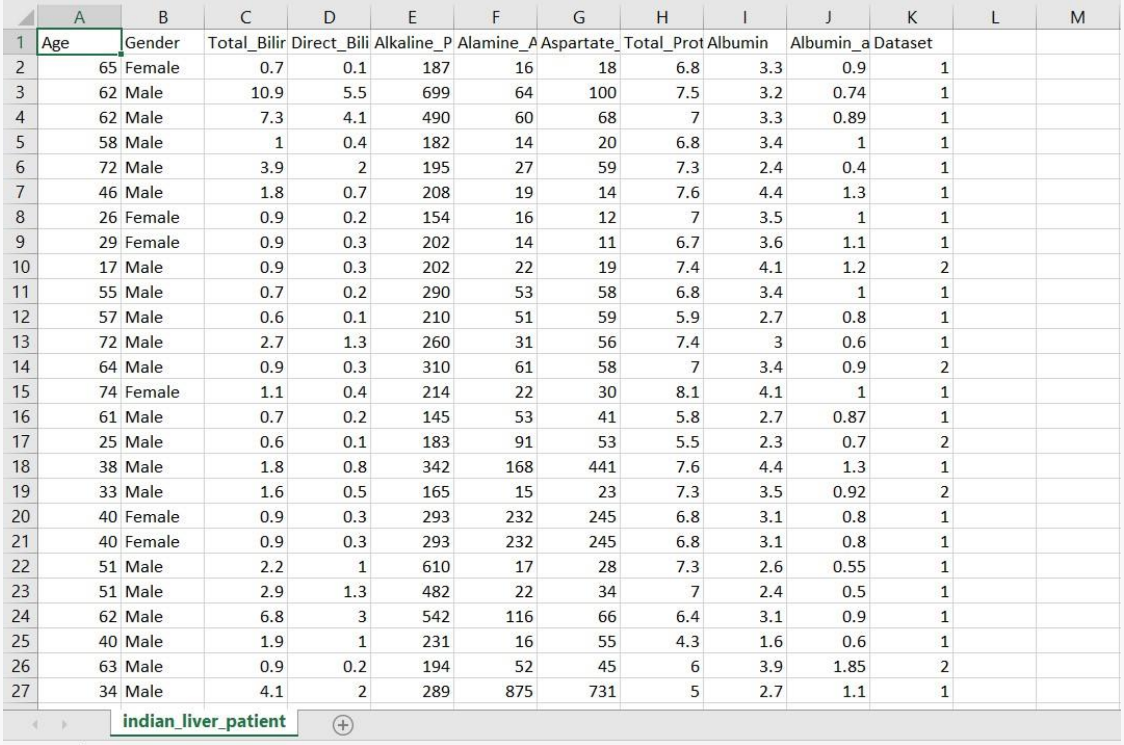
In 2022 at IEEE 8th conference , authors proposed Liver Disease Prediction Using Ensemble Technique. They suggested a system in which ensemble techniques such as random forest, xgboost, and gradient boost are merged to improve accuracy. Ensemble learning is a technique in machine learning where multiple models are combined to improve the overall performance of the system. The idea is that by combining multiple models, we can reduce the risk of making an incorrect prediction or classification due to the errors of a single model and it has been shown to improve the accuracy and robustness of machine learning models in many applications.[3]

3. DATASETS USED

The data is gathered from the Kaggle repository, and it forecasts liver illness based on the characteristics provided. The dataset contains eleven characteristics that indicate liver disease. The characteristics are described below, with the variables ordered by datatype.
The collection is constructed using integer datatypes. The dataset in this study includes attributes like Age, Gender, Total Bilirubin, Direct Bilirubin, Total Proteins, Albumin, Albumin and Globulin Ratio (A/G ratio). The collection includes 583 cases, 416 of which are positive for liver disease and 167 of which are negative. If the class designation has a class value of '1,' it means the patient has liver disease; if the class value is '2,' it means the patient does not have liver disease.[4]

Table 1: Attribute Description

Attributes	**Possible Value**
Age of the patient	Numeric
Gender of the patient	Nominal
Total Bilirubin	Numeric
Direct Bilirubin	Numeric
Alkphos Alkaline Phospotase	Numeric
Sgpt (Alamine Aminotransferase)	Numeric
Sgot (Aspartate aminotransferase)	Numeric
Total proteins	Numeric
Albumin	Numeric
Albumin and Globulin Ratio	Numeric
Dataset	Numeric



	A	B	C	D	E	F	G	H	I	J	K
1	Age	Gender	Total_Bilir	Direct_Bili	Alkaline_P	Alamine_A	Aspartate_	Total_Prot	Albumin	Albumin_a	Dataset
2	65	Female	0.7	0.1	187	16	18	6.8	3.3	0.9	1
3	62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
4	62	Male	7.3	4.1	490	60	68	7	3.3	0.89	1
5	58	Male	1	0.4	182	14	20	6.8	3.4	1	1
6	72	Male	3.9	2	195	27	59	7.3	2.4	0.4	1
7	46	Male	1.8	0.7	208	19	14	7.6	4.4	1.3	1
8	26	Female	0.9	0.2	154	16	12	7	3.5	1	1
9	29	Female	0.9	0.3	202	14	11	6.7	3.6	1.1	1
10	17	Male	0.9	0.3	202	22	19	7.4	4.1	1.2	2
11	55	Male	0.7	0.2	290	53	58	6.8	3.4	1	1
12	57	Male	0.6	0.1	210	51	59	5.9	2.7	0.8	1
13	72	Male	2.7	1.3	260	31	56	7.4	3	0.6	1
14	64	Male	0.9	0.3	310	61	58	7	3.4	0.9	2
15	74	Female	1.1	0.4	214	22	30	8.1	4.1	1	1
16	61	Male	0.7	0.2	145	53	41	5.8	2.7	0.87	1
17	25	Male	0.6	0.1	183	91	53	5.5	2.3	0.7	2
18	38	Male	1.8	0.8	342	168	441	7.6	4.4	1.3	1
19	33	Male	1.6	0.5	165	15	23	7.3	3.5	0.92	2
20	40	Female	0.9	0.3	293	232	245	6.8	3.1	0.8	1
21	40	Female	0.9	0.3	293	232	245	6.8	3.1	0.8	1
22	51	Male	2.2	1	610	17	28	7.3	2.6	0.55	1
23	51	Male	2.9	1.3	482	22	34	7	2.4	0.5	1
24	62	Male	6.8	3	542	116	66	6.4	3.1	0.9	1
25	40	Male	1.9	1	231	16	55	4.3	1.6	0.6	1
26	63	Male	0.9	0.2	194	52	45	6	3.9	1.85	2
27	34	Male	4.1	2	289	875	731	5	2.7	1.1	1

Fig.1 Dataset for Liver Disease Prediction

4. PROPOSED MODEL

Ensemble-based systems have become increasingly popular in various fields due to their ability to improve the accuracy and robustness of machine learning models. By combining multiple classifiers, an ensemble system can effectively mitigate the weaknesses of individual models and produce more accurate and reliable predictions However, the concept of using multiple experts to make decisions is not a new one. In fact, it is a practice that is deeply rooted in our daily lives, as highlighted in the examples mentioned in the paragraph. By seeking the opinions and expertise of multiple individuals, we increase the chances of making a better decision and reducing the likelihood of making a poor one. The psychological backdrop of this approach is based on the concept of diversity and the wisdom of crowds. By combining the opinions of multiple experts, we can leverage their individual strengths and knowledge, while also reducing the impact of individual biases and errors. This approach has been shown to be effective in a wide range of scenarios, from medical decision-making to financial forecasting and even political elections.[5]

In summary, the use of ensemble-based systems is a statistically sound approach that can improve the accuracy and robustness of machine learning models. However, the psychological backdrop of this approach, which is based on the concept of diversity and the wisdom of crowds, highlights the practical and everyday relevance of this approach in our daily lives. By seeking the opinions and expertise of multiple individuals, we can make better decisions and minimize the impact of individual biases and errors.
Ensemble learning techniques can be applied to image classification tasks to improve the accuracy and robustness of machine learning models. Here are some examples of ensemble learning techniques that can be used for image classification:

1. Bagging: Bagging stands for bootstrap aggregating, and it involves training multiple models on different subsets of the training data. In image classification, this means dividing the training data into different subsets and training multiple models on each subset. The predictions of these models are then combined to make a final prediction. This approach can help reduce overfitting and improve the accuracy of the model.
2. Boosting: Boosting involves training multiple models sequentially, with each model focusing on the examples that the previous model misclassified. In image classification, this means training multiple models on the training data, with each model focusing on the examples that the previous model misclassified. The predictions of these models are then combined to make a final prediction. This approach can help improve the accuracy of the model, particularly for difficult examples.
3. Stacking: Stacking involves training multiple models on the training data and then using the predictions of these models as input to a final model. In image classification, this means training multiple models on the training data and then using their predictions as input to a final model. This approach can help improve the accuracy of the model and reduce overfitting.

4.1 NAÏVE BAYES ALGORITHM

Naïve Bayes is a simple but powerful supervised learning algorithm that is used for classification tasks. It is based on Bayes theorem, which describes the probability of an event based on prior knowledge of conditions that might be related to the event. The Naïve Bayes classifier assumes that the presence or absence of a particular feature is independent of the presence or absence of any other feature, hence the term "naïve"[8]. The algorithm begins by determining the prior likelihood of each class based on the training data. Then, for a given input, it calculates the posterior probability of each class based on the prior probability and the likelihood of the input given each class. Finally, the algorithm assigns the input to the class with the highest posterior probability.The likelihood of the input given each class is calculated by multiplying the probabilities of each feature given the class.

The independence assumption of the Naïve Bayes classifier means that the joint probability of the features given the class can be factorized into the product of the probabilities of each individual feature given the class. Naïve Bayes classifiers are simple and efficient, and they work well in many real-world situations. They are particularly useful for text classification tasks, such as spam filtering and sentiment analysis. However, the strong independence assumption can be a limitation in some cases where the features are actually correlated.

In summary, Naïve Bayes is a supervised learning algorithm used for classification tasks, based on Bayes theorem and the assumption of independence between features. It calculates the posterior probability of each class given the input and assigns the input to the class with the highest posterior probability. Naïve Bayes classifiers are simple and efficient, and they work well in many real-world situations.

$$\mathbf{P(A/B)} = \frac{\mathbf{P(B/A)P(A)}}{\mathbf{P(B)}}$$

- P(A/B) is the posterior probability of class given predictor
- P(A) is the prior probability of class
- P(B/A) is the likelihood which is the probability of the predictor given class
- P(B) is the prior probability of the predictor

Using Bayesian probability terminology, the above equation can be written as [8]

$$\text{Posterior} = \frac{\text{Prior x likelihood}}{\text{Evidence}}$$

There are 3 types of Naïve Bayes Classifiers. They are :

- Multinomial Naïve Bayes Classifier
- Bernoulli Naïve Bayes Classifier
- Guassian Naïve Bayes Classifier

Here in this model we have used Bernoulli Naïve Bayes classifier which is useful for binary distribution where the output label may be present or absent.

Bernoulli Naïve Bayes Classifier

The Bernoulli Naïve Bayes classifier is a variant of the Naïve Bayes classifier used in machine learning. It is specifically designed for datasets where the features are binary-valued (i.e., they can take on only two values, such as yes or no), making it a useful algorithm for tasks like text classification. The Bernoulli Naïve Bayes classifier assumes that each feature is a binaryvalued variable that follows a Bernoulli distribution. In other words, it assumes that each feature is either present or absent in a given example. Based on this assumption, the classifier calculates the likelihood of each class given the presence or absence of each feature in the input.

Like other variants of the Naïve Bayes classifier, the Bernoulli Naïve Bayes classifier uses Bayes' theorem to compute the probability of each class given the input. The classifier starts by calculating the prior probability of each class based on the frequency of each class in the training data. Then, for a given input, it calculates the posterior probability of each class based on the prior probability and the likelihood of the input given each class. To use the Bernoulli Naïve Bayes classifier, the input data must be represented as binary-valued feature vectors. Each feature in the vector represents a binary-valued variable that is either present or absent in the input example. The classifier then uses these feature vectors to calculate the posterior probability of each class and assigns the input to the class with the highest posterior probability. It is a reduced model in which counting the number of words is less essential, and Bernoulli may produce superior results. It only works for binary values and produce results that are computationally better than many traditional algorithms.

DECISION STUMP:

A decision stump is a simple machine-learning model that consists of a one-level decision tree. Unlike full decision trees, which can have many levels and internal nodes, a decision stump has only a single internal node (i.e., the root) that splits the data based on the value of a single input feature. The root node of the decision stump makes a decision based on a threshold value for the input feature, and the two branches correspond to the two possible outcomes (e.g., yes or no).If the input feature is discrete (i.e., it can take on a limited number of values), then the decision stump consists of a single interior node. If the feature is numerical, the decision stump may be more complex and involve multiple interior nodes. The leaf nodes of a decision stump contain the predicted class name, while the non-leaf nodes are decision nodes that test the value of the input feature. Each branch of a decision node corresponds to a possible value of the input feature, and the input data is partitioned accordingly.

ADABOOST:

A common method for solving binary categorization issues is boosting. By transforming a number of poor learners into strong learners, these programmes increase prediction power. Boosting algorithms work on the premise that we first create a model using the training dataset, and then we create a second model to fix any mistakes in the first model. The procedure is repeated up until the mistakes are reduced and the dataset's predictions are accurate. Similar to this, boosting algorithms combine various models to produce the desired result.[7] Adaboost, also known as Adaptive Boosting, is a machine learning strategy used as an ensemble approach. It is optimal to improve decision tree efficiency for binary classification. It is best used with weak pupils and is used for categorization. [7]

4.2 SYSTEM ARCHITECTURE

For execution, the Data is loaded and prepared.

After feature selection, we run the data via a machine learning algorithm.

After training, the model can be used to make additional predictions.

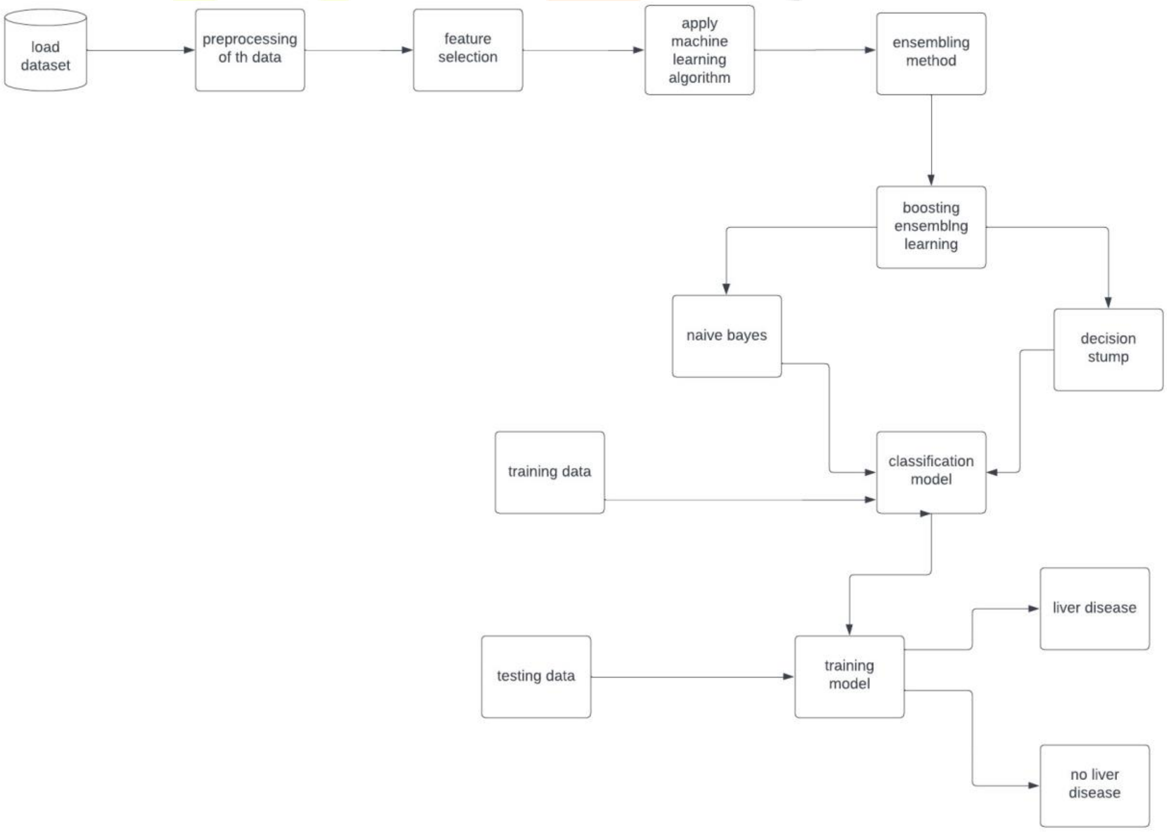


Fig: System architecture for liver disease prediction using Naïve Bayes and Decision stump

5. TESTING AND RESULT

The dataset is imported into a Jupyter notebook, and the data is preprocessed using the Individually Naive Bayes and Decision Stump algorithms. Out of 583 cases, 507 are trained at random, and the remaining 76 are evaluated. Later, to improve the precision of the model assembly method is used. The suggested work's findings were acquired by employing the Naive Bayes and Decision Stump algorithms to an Indian Liver Disease dataset, yielding accuracy of 72.3% and 76.3%, respectively. After combining those two methods, the machine learning model achieved a precision of 94.7%.

6. CONCLUSION

Classification is a supervised machine learning technique that is primarily used in the healthcare sector for medical diagnosis and disease prediction. This model was created by combining two algorithms: Naive Bayes and Decision Stump. Individuals' health will profit from the use of Decision stump in predicting liver illness. Furthermore, the assembling technique outperforms separate methods in terms of accuracy. However, in the future, we will gather the most current data from different regions around the globe and train the model for detecting liver disease. The findings of this research will motivate us to continue creating advanced versions of these types of machine learning algorithms.

7. REFERENCES

1. *S. Aneeshkumar and C.J. Venkateswaran, "Estimating the Surveillance of Liver Disorder using Classification Algorithms", International Journal of Computer Applications (0975 –8887) , Vol. 57, no. 6, (2012).*
2. *S.Dhamodharan, "Liver Disease Prediction Using Bayesian Classification", 4th National Conference on Advanced Computing, Applications & Technologies, Special Issue, May 2014.*
3. *Sai Rohith Tanaku , Addike Ajay Kumar and co-authors 2022 'Liver disease using Prediction using Ensemble Technique' 25 – 26 March 2022*
4. *Indian Liver Patient Records (2018) [Version 1] retrieved from [https://www.kaggle.com/datasets/uciml/indian-liver-patient-records]*
5. *Aishwarya Singh (2018) Analytics vidya December 19, 2022 Ensembling models [https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemblemodels/]*
6. *Tobias Schlagenhauf Feb 17, 2022 python -course.eu Boosting algorithm in python [https://python-course.eu/machine-learning/boosting-algorithm-in-python.php]*
7. *James Ajeeth Dec 2020 KDnuggets Implementing Adaboost Algorithm from Scratch [https://www.kdnuggets.com/2020/12/implementing-adaboost-algorithm-fromscratch.html]*
8. *Nagesh Singh Chauhan (April 8, 2022) KDnuggets Naïve Bayes Algorithm [https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html]*