Journal of Theoretical and Applied Information Technology
30th June 2022. Vol.100. No 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SOFT VOTING MACHINE LEARNING CLASSIFICATION MODEL TO PREDICT AND EXPOSE LIVER DISORDER FOR HUMAN PATIENTS

MOHAMMAD A. ALSHARAIAH *[1] , LAITH H. BANIATA [2], OMAR AL ADWAN [3] , ORIEB ABU ALGHANAM [4] ,AHMAD ADEL ABU SHAREHA[5], MOSLEH ABU ALHAJ [6], QAIS ABDALLAH SHARAYAH [7] ,MOHAMMAD BANIATA[8]

[1,3,4,5,6]Al-Ahliyya Amman University, Amman- Jordan
[2]Gachon University, South Korea
[7]Jordanian Royal Medical Services, Amman-Jordan
[8] Ubion , South Korea
Email: [1] m.sharaiah@ammanu.edu.jo

* Corresponding author: m.sharaiah@ammanu.edu.jo

ABSTRACT

The liver is the most significant organ in the human body since it handles a significant role in food digestion and progression in our body. Mainly it takes an essential part in enzyme activation, fat metabolism, bile synthesis, vitamin, glycogen, and mineral storage. Depending on the role it controls, it has a sophisticated accidental of coming in contact with harmful creation that goes inside the body. Hence, the diagnosis of liver disorder has been subjective at best, based on subjective approaches. Liver disorders are challenging to detect, and as a result, they are regularly overlooked in the early stages due to a lack of precise symptoms. Hyperbilirubinemia is one of the most substantial signs of most liver illnesses, and it can be demanding to differentiate early on. However, in most of cases, this isn't certain, and the ability to detect and confirm the presence of liver disease lead to a better understanding of enzyme levels. The prediction of liver illnesses has been done using a variety of machine learning techniques. In this investigation, the recommended ensemble soft voting classifier offers binary classification and utilize the ensemble of three machine learning algorithms: Decision Tree, Support Vector Machine, and Naive bayes classifiers to predict and expose liver disease by Binary Classification of the dataset into two particular types of patients with or without liver disease (patient suffering liver sickness or not). The unbalanced dataset comprises materials about human patient attributes such as Gender, Age, Alanine, Total Bilirubin, Aminotransferase, Aspartate Aminotransferase, Direct Bilirubin, Albumin, Alkaline Phosphatase, Globulin Ratio and the Result and Total Proteins Albumin. Furthermore, the accuracy and various error calculations of the predictions from the aforementioned algorithms are analyzed to recognize and identify the best- convenient algorithm.

Keywords: *Liver Disorders, Machine Learning Techniques, Unbalanced Data Set, Classification, Liver Hyperbilirubinemia*

1. INTRODUCTION

The liver, is apart in organism body called an Exocrine Gland, is found in the right side of the stomach, below the diaphragm. It is in charge for a variety of important life processes, including bile production for digestion, blood cleansing, blood toxicity management, bilirubin clearance, body metabolism, and the conversion of hazardous ammonia to urea. Full of fat Liver Disease is initiated by the accretion of fat in the liver. Fatty Liver Disease, which is caused by the accumulation of fat in the liver, is a very common condition in various world countries. For instance, in India with over 10 million cases reported each year. Testing is essential for diagnosis due to the shortage of adverse effects. Hepatic fibrosis and associated end-stage cirrhosis are a growing problem around the world. Cirrhosis is the permanent consequence of fibrosis scarring, with interrelating bands of fibrosis tissue swapping normal liver architecture. The hepatitis B, hepatitis C, and excessive alcohol consumption are the most prevalent etiological features that chief to cirrhosis. [1]. Chronic HCV contagion is typically a slow, progressive infection that might cause few or no indications for several years afterward infection. Certain patients progress chronic infection and suffer no substantial liver injury, while others

progress quickly to liver cirrhosis and might mature hepatocellular carcinoma [2]. Patients through chronic liver illness are at advanced danger of emerging hepatocellular carcinoma and must be observed on a frequent basis for initial detection. Chronic HCV infection is the major reason of cirrhosis and hepatocellular carcinoma (HCC). Herein, alpha fetoprotein intensities might be raised. Hepatocellular carcinoma is becoming further mutual, and this tendency is predicted to endure for numerous years [3]. According to present research, the mutual of HCC patients advanced the disease as an outcome of a build-up of genetic irregularities, which were most probable instigated through external etiological issues such as HBV and HCV contagions. [4]. These hazard factors be able to cause DNA sequence variations and injury, such as the p53 modification caused by aflatoxin besides DNA damage initiated thru the HBV genome interruption. [5]. Averting hepatitis virus infection and eliminating hepatitis virus in long-lasting hepatitis patients is critical in stopping liver cancer.

However, Artificial Intelligence (AI) is a category of computer cognitive that is authorized by the capability of computer plans to comprehend, obtain knowledge, and then use that knowledge in various fields [6]. Human awareness is currently utilized in practically every sphere of application, and it is made up of numerous components, such as Deep and Machine Learning. These arenas can be subdivided into a variety of sections. Machine Learning is a subclass of Artificial Intelligence that comprises procedures to make the machine from previous data in demand to allow it to make decisions in a specific condition may decrease the load on doctors. Besides, Machine Learning is divided into two subclasses: supervised and unsupervised machine learning. The Supervised Machine Learning is a learning method in which the machine is trained through a labeled dataset. This is called the training dataset, and it contains the labels that agree to the exact responses. In addition, the machine's performance is validated utilizing the testing dataset after it has been trained. Support vector machine (SVM), random tree technique and naïve bayse classifier are some illustrations of Supervised Machine Learning. When the dataset does not require a label, the Unsupervised Machine learning method can be applied. Instead, the machine learns by building clusters out of previously unknown patterns in the given information [7]. Another method can be useful, Semi-Supervised Learning method is a learning strategy that combines two methodologies, notably Supervised Learning and Unsupervised Learning, in that certain of the data is categorized with labeled and the other is not, and the training and testing sets are not stated obviously. The research validates the enhanced method by ensemble of three machine learning algorithms through soft voting classifier.

1.1 motivation

The main motivation of this paper is the challenge of diagnosis and detect the liver disorder in early stages due to the lack of precis symptoms. As discussed earlier in this section, various classification approaches have been reported in the literature however, most of these techniques are not viable for the liver disorder classification

1.2 Contribution

To overcome the aforementioned challenge, a soft voting approach that uses an assemble of three algorithms is proposed. To achieve the main goals of this research study, outcomes have been investigated with liver disease dataset. In this research article, the research objective and the major contributions are presented as follows.

1. An collaborative of machine learning algorithms: Decision Tree, Naïve Bayes and SVM, with soft voting classifier have been proposed. The suggested methodology binary classifies the liver disease data within infected and un-infected classes.
2. Experiments have been applied on liver disease dataset.
3. Precision, Accuracy, recall, AUC curve, F1-score have been occupied as the assessment criteria for testing the robustness of recommended methodology.
4. Employing the soft voting machine learning classifier for the liver disorder.
5. Exploring the efficiency for the soft voting machine learning classifier over other classical machine learning methods.

1.3 paper outline

The remainder of the article is organized as follows. In Section II, literature review is presented. Section III encloses the utilized method and model architecture and implementation. Section IV encloses the results and investigation. Section V presents the conclusions and future scope by section.

2. RELATED WORKS

Machine Learning has exposed promising and efficient results in several fields [8] [9], especially the field of medicinal science to assist medical

doctors in the procedure of finding and decrease their load. As considered within several latest research works, [10] Machine learning algorithms such as Random Tree, LMP, J48, Random Forest, Hoeffding and Decision Stump which are interrelated to Decision Tree are utilized in classifying the available dataset. In addition, concerning to the performance of each algorithm, the Precision, runtime, recall, mean absolute error, and accuracy are utilized as performance measurements, and the findings demonstrate that the Decision Base has the extreme accuracy. [11] has recommended the utilize of classification algorithms specifically, KNN, ANN, Logistic Regression and SVM with back propagation which includes of 10 input neurons layers and the examination demonstrations that ANN is sensibly efficient. In an investigation, genetic microarrays in addition to the neural classifications has been recommended by [12] [13] for the estimates and prediction, this has stated atomic science method to increase the predictive ability. The results accomplished from the investigations in [13] have definite that Random Forest calculation isn't appropriate subsequently the concern of over fitting and recommendations headed for the utilize of oversampling approaches have been arranged to address this substance. Additionally, [14] claims that while both K-NN and Logistic Regression algorithms provide similar (exactness) precision, performance in therapeutic research can be distinguished by sensitivity. [15] Proposes that the abstinence paradigm may be utilized to avoid inaccurate categorization by ensuring that no outcome is formed when the model is unsure about its prediction. Besides, the research in [16] reveal that biochemical indicators have a huge impact on the positive anticipation of different stages of liver fibrosis and the level of Hyaluronic Acid (HA). Moreover, [17] has offered that in distinction with Resilient Backpropagation Neural Network (Rprop) and Stochastic Average Gradient (SAG) proceeding the dataset with text, and Convolutional Neural Network (CNN) for image dataset, the peak precision is realized by CNN. [18] Used a variety of algorithms, including Random, Forest K-Means, K-NN, Naive Bayes, and C 5.0, to illustration that adaptive boosting increases C 5.0's performance. However, [19] has used a variety of thorough and exclusive parameters to select the data needed to train the model, ensuring that unpredictably structured data does not affect classification. In addition, [20] in his study advocated using unsupervised learning approaches such as affinity Propagation, K-Means and DBSCAN to detect the Silhouette coefficient as serious for creation a performance assortment.

Data mining methods have extensively been utilized for the prediction of several illnesses. In [21], the authors define the use of classification algorithms such as Bayes Algorithm, Decision tree algorithm, and Rule based Algorithm intended for diabetes illness prediction and they are deliberated common classification technique at the time. To advance the efficiency of classification algorithm the feature extraction was utilized. In [22], several classification algorithms specifically K – Nearest Neighbour, Support Vector Machine and Logistic Regression have been employed for the prediction of liver infection. On rely of sensitivity, the last classification algorithm has showed to be further suitable aimed at the prediction of the illness. As well, in [23], the authors clarify different unsupervised classification methods for the classification of the dataset. The three methods utilized in the paper are Affinity Propagation, DBSCAN and K-means. The recommended technique is separated into three stages explicitly analysis, prediction and comparison. The analysis is being prepared using K-means, Affinity Propagation and DBSCAN. The data set employed encloses several levels of enzymes present in the liver system. In demand to discovery the greatest, performance of the procedures is implemented and is intended using essentially Silhouette Coefficient. This factor defines accurateness and number of cluster which defines complexity. Lastly, K-means is discovered to be the optimum technique in association to other. The paper more points to future work in purpose of other illnesses such as lung, heart, and brain. Scientists utilized the collaborative technique in which numerous single models are joined to provide improved prediction outcomes. In 2014, Vijayan et al [30] used several data mining methods designed for diabetes mellitus . In 2017, the author shared the significance of AdaBoost and bagging procedures of machine learning, the usage of 48 as the basis for diabetes prediction. It excitingly classifies diabetic and non-diabetic patients depend on specific factors such as risk factors of diabetes. It was approved that the AdaBoost learning algorithm surpasses than J48 and bagging algorithm [31]. Knowler & Johannes, Smith, Everhart, Dickson, [32] offered a neural network with ADAP algorithm to shape an associative model in which they arbitrarily choose data for training and the exactness succeeded was 76%. Quinlan [33] engaged a C4.5 learning model and the model achieved fine with an accurateness of 71.1%. Furthermore, J48, Radial basis function Naive Bayes, Artificial neural network had been

hired for diabetes kind 2 diagnosis. Naive Bayes reached an accurateness of 76.95% and outperformed results of J48, RBF, with accurateness of 76.52%, 74.34% individually (Nai-Arun & Moungmai [34]. In 2015, Nongyao et al. Offered a model that meditate the danger of diabetes. The approach used four several machine learning algorithms for the classification persistence: LR, ANN, NB and DT. Bagging and boosting have been used for invention in outcomes. The trials prove that the random forest algorithm outperformed as matched to other algorithms Soltani & Jafarian [35]. For instance, Sahan et al. hired a 10- fold cross-validation method through a weighted artificial immune system and acquired a prediction accurateness of 75.87%, Sahan, Kodaz & Gunes [36]. Anand et al. appoint the CART model and the algorithm an accomplished accurateness of 75% Anand & Shakti, [37]. Also, Rani and Jyothi [38] presented collaborative algorithms that use the filtered classifier, KNN, ANN, NB, zeroR, simple cart, J48, and cv parameter selection regarding to the classification. However, the presented methodology attained an accurateness of 77.01%. In addition, Li L. proposed a methodology with ANN, SVM, Naïve Bayes and a weighted based investigation for the classification Li. [39]. Finally, Bashir et al. proposed a collaborative model of CART and C4.5 which reached an accurateness of 76.5% (Bashir, Qamar, Khan & Javed, [40]) .

3. PROPOSED WORK AND OVERVIEW OF SYSTEM ARCHITECTURE

Incision the dataset inside two sets, training and testing sets is the major step in the implementation. The model will be trained then applied to the testing set, and the results will be assessed utilizing performance metrics. The design for the model has been shown in Figure.1 and Figure.2. In addition, the implementation for the proposed model has be done by employing python programming language. For instance, data analysis, data preparation, data training and testing have been proceeded by special python codes and Scikit-learn - Machine Learning Library.

3.1 Data Collection

The data chosen is the crucial and substantial step in machine learning algorithm implementation. In this article we have selected a secondary data such as the Indian Liver Patient Dataset, which is part of the UCI Machine Learning Fountain, was utilized to compile the essential data for this study (ILPD) [23].

3.2 Data Preprocessing

In the machine learning process, data exploration is the first step in evaluating the algorithm. Therefore, in this study, we engaged the CSV file layout (Comma Separated Values). Pre-processing the specified dataset is critical for the python application's algorithm implementation. For instance, Data discretization, resampling, data normalization and attribute choices are the preprocessing instrument. Filtering the dataset is another term for pre-processing. All of the attributes were discretized in this paper. The discretization will group the different values and transform all attribute values from integer to nominal values. However, the numeric toward nominal values filter was also used after the discretization was completed. Actually, the available raw data acquired is noisy, fragmentary, and inconsistent, it is not appropriate for immediate use. As a result, cleaning is required to convert the raw data to useful data for additional processing. Consequently, this procedure involves coping with null values and swapping them with computed mean values.

Journal of Theoretical and Applied Information Technology
30th June 2022. Vol.100. No 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

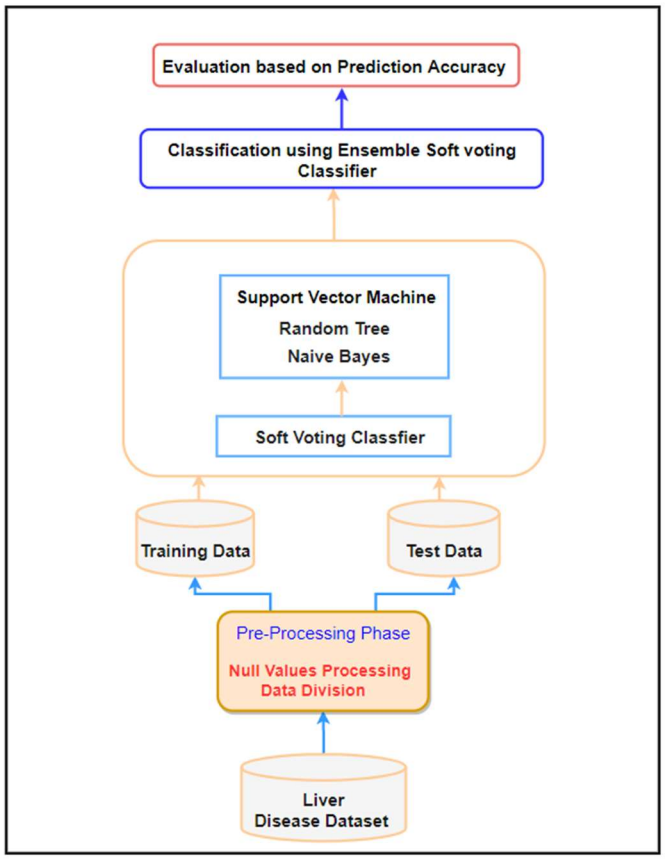

Figure1: The Proposed Model

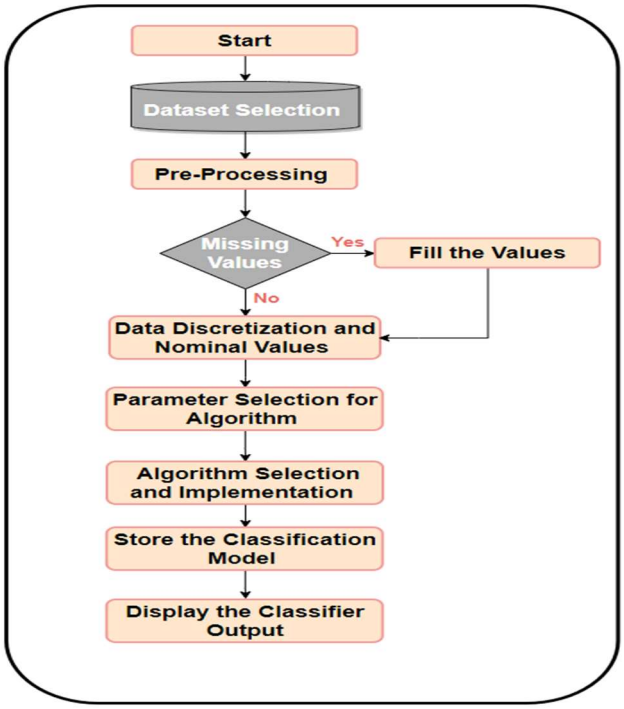

Figure2: The Architecture Of Data Flow.

3.3 Data analysis

The openly obtainable Indian liver patient dataset from University of California Irvine machine learning dataset source [24] is engaged for this work. This data is gathered from patients of specific area such as north-east Andhra Pradesh, India. It encloses 583 samples including 416 unhealthy liver samples and remaining with number of 167 non-liver diseased samples. Its data is tabularized through 10 input attributes and only single output class attribute. The attribute information details of the dataset are specified in Table 1.

Table 1: List Of Attributes

No	Attribute name	Attribute description
1	Age	Age of the patient
2	Gender	Gender of the patient
3	Tot_bilirubin	Total Bilirubin
4	Direct_ bilirubin	Direct Bilirubin
5	Tot-proteins	Total proteins present in patient
6	Albumin	Albumine amount of the patient
7	Ag_ratio	Albumine and Globuline ratio

8	Sgpt	Alamine Aminotransferase
9	Sgot	Aspartate Aminotransferase
10	Alkphos	Alkaline Phosphatase
11	Is_patient	Whether the data is belongs to liver disease patient or not

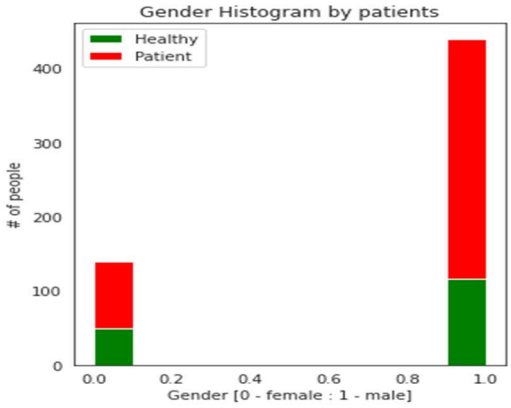

Figure 3: Data Set Analysis

Figure.3 illustrates the dataset statistics after completing a few pre-processing stages using the python functions. Figure.3 reveals that among the 514 patients, 167 patients were classified as healthy cases. 40 cases of 167 patients are female patients and the rest are classified as male patients. Also, Figure.3 shows that among the 514 patients, 416 cases were classified as infected cases. 90 cases of 416 are female patients and, the rest were classified as male patients. This statistic is an indication that the male gender is more exposed than the female gender to be infected by liver diseases due to many various factors mentioned in the introduction section.
Figure.3. Data set analysis

3.4 Model Architecture

The major purpose of this study is to train the model to accurately categorize a specified dataset of patient parameters into the classes of either Diseased or Not Diseased. The classification algorithms applied are, The Decision Tree Support Vector Machine (SVM) and Naïve bayes. However, On the basis of several performance measures listed below, the outcomes of the performance of these algorithms on the dataset will be matched

.

3.4.1 Support vector machine (SVM)

Classification methods are broadly used in several medical implementations. Classification purposes to shape an active model for predicting class labels of unidentified data. The model is constructed on the training data, which involves of data points selected from input data domain and their class labels. A Support Vector Machine (SVM) splits the data into binary categories of accomplishment classification and creating an N-dimensional hyper plane. These models are closely associated to traditional multilayer perceptron neural networks. A support vector machine builds a hyper plane or set of hyper planes in a high- or infinite-dimensional space. A proper separation is realized through the hyper plane that has the major distance to the closest training data point of any class (so-named functional margin), as in common the larger the margin the lower the generality error of the classifier. There are an another training method for radial basis function, polynomial and multi-layer perceptron classifiers in which the weights of the network are initiated by explaining a quadratic programming problematic by linear constraints, slightly than by solving a non-convex. However, the unconstrained minimization problematic as in regular neural network training [7], there are numerous likely kernel functions and the most mutual kernel are: Linear, sigmoid, polynomial and radial basis function (RBF). In this research paper we occupation linear kernel function which shows in equation .1:

$$K(Xi,Xj)=xiTxj \qquad (1)$$

relying on the kernel type we select the kernel parameters have to be fixed. Which kernel type achieves optimum, based on the application and be able to be configured through using cross-validation. In the literature, SVM, a predictor variable which is termed an attribute and a converted attribute that is used to describe the hyper plane is entitled a feature [25]. At this time, selecting the best appropriate representation can be taken as feature selection. A set of features that defines one case is termed a vector. The aim of this modeling is to discovery the optimum hyperplane which splits clusters of vector. The vectors close to the hyper plane are the support vectors [26] as in Figure.4.

Journal of Theoretical and Applied Information Technology
30th June 2022. Vol.100. No 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

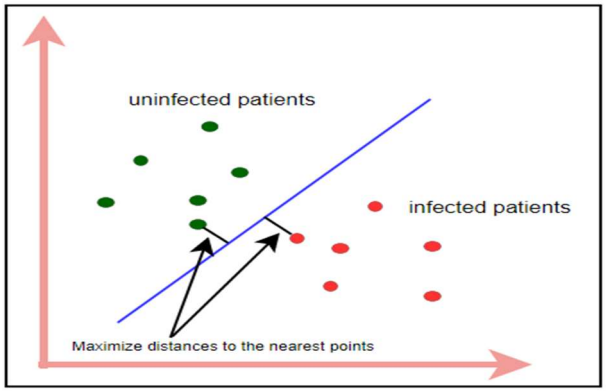

Figure. 4. Support Vector Machine

3.4.2 Decision tree

This classifier is a kind of ensemble classifier. It utilizes decision tree models to acquire improved prediction consequences. It develops various trees and a bootstrap method is engaged to every tree from the set of training data. In classification, procedure input is provided to every tree present in the forest, and formerly, |every tree votes separately for that class. In the conclusion, the RF chooses the class, which has become the highest number of votes. Decision trees are the construction blocks of a random forest algorithm. Precisely, a decision tree is a decision support technique that customs a tree-like structure. A decision tree involves three components: decision nodes, leaf nodes, and a root node. A decision tree algorithm splits a training dataset into several branches, which auxiliary separate into other branches. This order remains until a leaf node is attained. The leaf node will be not be able to segregated more. The nodes in the decision tree signify attributes that are engaged for predicting the result. Decision nodes afford a link to the leaves. The illustration in Figure.5 shows the three kinds of nodes in a decision tree.

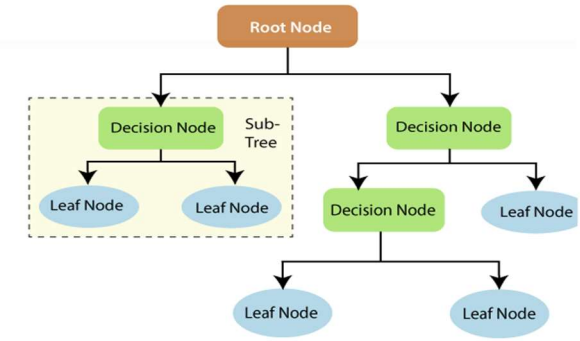

Figure.5. Decision Tree

3.4.3 The naive Bayes

The naive Bayes classifier is a straightforward "probabilistic classifier" established on Bayes' theorem and strong (nave) independence conventions among the features [27] They are a unique of the most elementary Bayesian network models (McCallum, 2019). However, when joint with kernel density approximation, they be able to accomplish higher levels of accurateness. [28] Classifiers are scalable, demanding a set of parameters that is relative to the amount of variables (features/predictors) in a learning matter. By judging a closed-form expression, maximum-likelihood training can be done. [6] [29]. It follows a hypothesis that the predictors are self-regulating. The existence of a specific feature or attribute in a class is not associated to the presence of any further feature or attribute can be considered as the elementary conjecture of Naive Bayes classifier.
Equation.2 is the formulation of the well-known Bayes' Theorem and provides the probability of incidence of an action after incidence of an action. At this point, P(R) is the priori of R. If y is the class variable and X is the dependent feature vector.

$$P(R|S) = \frac{P(S|R)\ P(R)}{P(S)} \quad (2)$$

Equation.3 indications the usage of the proposition to our dataset. The Naive supposition can be assimilated and the algorithm applied for a dataset.

$$P(y|X) = \frac{P(X|y)\ P(y)}{P(X)} \quad (3)$$

3.4.4 Proposed Ensemble soft voting classifier:

This classifier is an encountered classifier for merging same or theoretically different machine learning models for prediction over common voting. A voting classifier hiring two kinds of voting techniques, hard and soft. In hard voting, the last prediction is completed through a majority vote in which the aggregator chooses the class prediction that derives again and again amongst the base models. On other hand, in soft voting, base models should have especial method names the Predict_proba method. The voting classifier offerings improved overall outcomes than other base models, as it associations the predictions of different

models. In the proposed model, SVM, Naive Bayes, and Decision Tree classifier have been collaborative. A soft voting classifier has been used which uses the predict_proba attribute column that provides the probability of every target variable [42]. Then it shuffles training data including the data points, and these data points are delivered to SVM, Naïve Bayes, and Decision Tree model. Every model computes individual prediction with voting aggregator and soft voting technique, also the majority voting is calculated which yields the final prediction. The algorithm for the proposed methodology has been demonstrated in Figure.1.

3.4.4 Performance metrics:

The authors analyzed the many parameters used to assess each algorithm's performance in terms of error and accuracy. The most popular method which used to analysis the effectively of the model are listed below:

Mean Squared Error (MSE): It characterizes the average of the squares of the variance among the estimated value and accurate value.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_l) \qquad (4)$$

Root Mean Square Error (RMSE): It is an amount of the quadratic mean of the change of observed and estimated values.

$$RMSE = \sqrt{\frac{\sum_{l=1}^{n}(predicted_i - Actual_i)}{n}} \qquad (5)$$

Accuracy: It is the proportion of instances that have been properly classified by the model.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \qquad (6)$$

4. RESULTS AND DISCUSSIONS

The metric values obtained by applying Naïve Bayes, Support Vector Machine (SVM), Decision Tree Algorithms respectively on the testing dataset have been exposed in Figure.6, 7 and 8. As it can be seen from Figure. 6, 7 and 8, the Decision Tree has outperformed the Naïve Bayes and the SVM algorithms and obtained a 0.75 accuracy value and 0.24 MSE value on the testing dataset. Also, the NV algorithm has achieved a very good results and it scored a 0.56 accuracy value and 0.43 MSE value on the testing dataset. The SVM algorithm outperformed the Naïve Bayes in classifying the patients whether they are healthy or infected with liver disease.

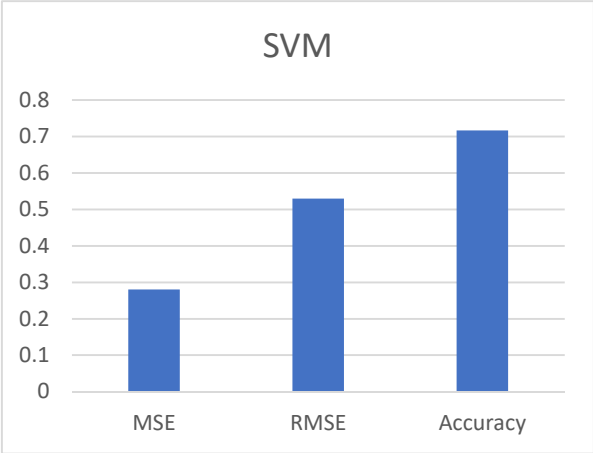

Figure.6: The Results Of SVM

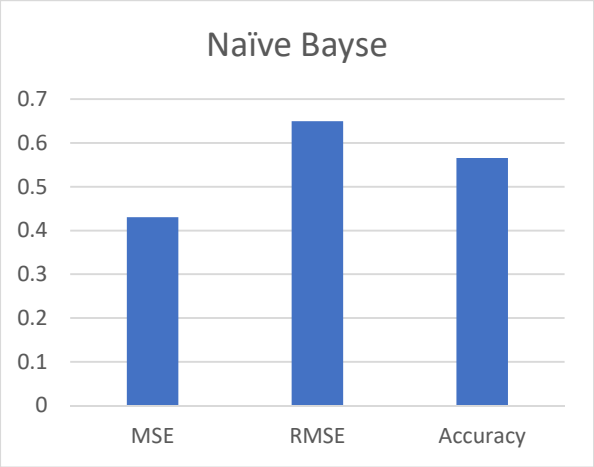

Figure7: The Results Of Naïve Bayse

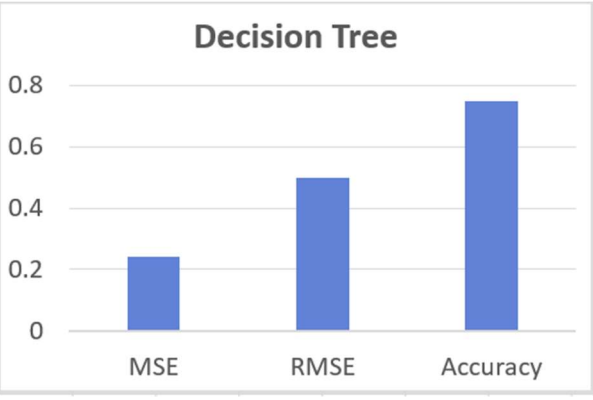

Figure 8: The Results Of Decision Tree

Journal of Theoretical and Applied Information Technology
30th June 2022. Vol.100. No 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

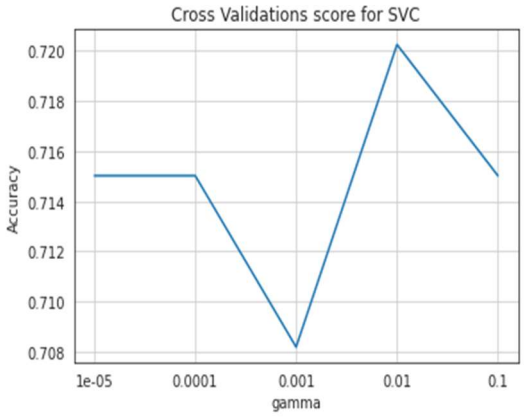

Figure 9: Cross Validation Score For SVC

Also, Figure. 9 illustrate the cross-validation score for SVM. It shows that when the SVM score a high accuracy value when gamma value is set to 0.01. The lowest accuracy is obtained by SVM when the gamma value is set to 0.001. As presented in the Table 2, SVM and NV have a closed performance but in terms of accurateness, SVM and SVC has given the best performance. Furthermore, the proposed soft voting approach was able to obtain remarkable accuracy score for the unbalanced liver dataset when compared with other algorithms as illustrated in Table 2. More importantly, results in table 2 shows that the performance of the proposed soft voting model is higher when compared to the KNN, Naïve Bayse, Random Forest and C5.0 that was proposed by Kumar [16]. Correspondingly, the values acquired for the numerous error metrics used, specifically Mean Squared Error (MSE), Root Mean Square Error (RMSE), and illustration that SVC has certain improved performance as accorded to the supplementary classification algorithms. Therefore, generally Soft voting classifier Algorithm has been exposed to accomplish best on this unbalanced dataset as shown in the Figure.10.

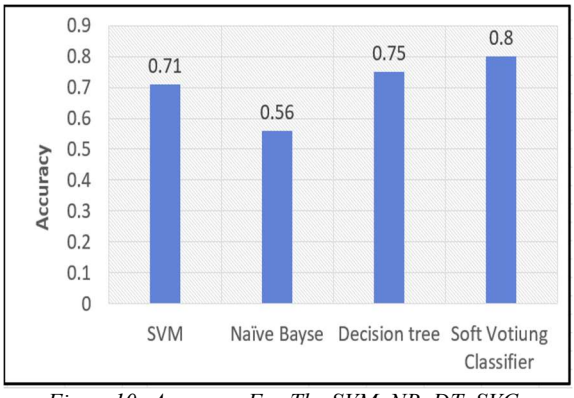

Figure10: Accuracy For The SVM, NB, DT, SVC

Table2 : Performance metrics results for SVM, Naïve Bayes and Decision Tree

Algorithm Name	MSE	RMSE	Accuracy
SVM	0.28	0.53	0.71
Naïve Bayse	0.43	0.65	0.56
Decision tree	0.24	0.49	0.75
[16], KNN	-	-	0.7068
[16], Naïve Bayse	-	-	0.6617
[16],Random forest	-	-	0.7218
[16], C5.0			0.7519
Soft Voting Classifier (SVC)	0.20	0.44	**0.80**

5. CONCLUSION

The main goal of this study is to examine the above-mentioned classification algorithms on an unbalanced dataset and assess their performance using parameters. The proposed ensemble soft voting classifier gives binary classification and uses the ensemble of three machine-learning algorithms: Decision Tree, Support Vector Machine, and Naive bayes classifiers to predict and expose liver disease. By training the proposed model on unbalanced dataset for liver disorder, the proposed model achieved a definite improvement of the classification performance. The results of the study suggest that the proposed model improved the classification accuracy score. In the current research work, the performance of the classification was significantly improved by adapting soft voting approach. This approach aids in identifying the algorithm that is most suited for this type of dataset with predefined attributes. The parameters for comparison are RMSE, MSE and Accuracy. In terms of Accuracy, Soft voting classifier has given the most precise predictions. Because liver illness isn't easily detectable in its early stages, medical specialists can use this model to accurately anticipate it. This study provides a helpful prediction model for the medical area people for the easy predictions. This model may be used for even larger datasets with more features in the future, causing the model to perform even better.

REFERENCES

[1] J. B. E. J. R. K. Golla, ""Liver disease: Current perspectives on medical and dental management"," Medical management update, pp. vol. 98 , No. 5, November 2004. .

[2] B. .. L. S. J. H. T.J. Liang, ""“Pathogenesis, natural history, treatment, and prevention of hepatitis C.”," Ann Intern Med, pp. pp132:296,vol.305, 2000.
[3] P. Johnson., ""“Hepatocellular carcinomaa: is current therapy really altering outcome," Gut, pp. , pp51:459, vol.62, 2002.
[4] E. M. H. A. S. M. S. Mabrouk, "“‘Statistical Approaches for Hepatocellular Carcinoma (HCC) Biomarker Discovery”," American Journal of Bioinformatics Research, pp. Vol. 2 No. 6, pp. 102-109, 2012.
[5] E. M. H. ,. A. S. M. S. Mabrouk, ""“Discrete Stationary Wavelet Transform of Array CGH Data on Hepatocellular Carcinoma’’," Journal of Bioinformatics and Intelligent Control, pp. vol.1,No 2, 2013.
[6] S. Russell and P. Norvig, "Artificial Intelligence: A Modern Approach(2nd ed.)," Prentice Hall, 2003.
[7] M. P. ,. N. B. V. B. V. Ramana, "A Critical Study of Selected Classification algorithms for Liver Disease Diagnosis," International Journal of Database Management Systems (IJDMS), pp. Vol.3, pp 111:114, 2011.
[8] F. A. Nazmun Nahar, ""Liver disease prediction by using different Decision tree techniques"," International Journal of Data Mining & Knowledge Management Process (IJDKP), pp. Vol.8, No.2, March, 2018.
[9] J. C. M. J. M. E. I. Joel Jacob, ""Diagnosis of Liver Disease Using Machine Learning Techniques"," International Research Journal of Engineering and Technology, 2018.
[10] J. L. R. D. Sumedh Sontakke, ""“Diagnosis of Liver Diseases using Machine Learning”," International Conference on Emerging Trends & Innovation in ICT (ICEI) , 2017.
[11] P. K. Shambel Kefelegn, ""Prediction and Analysis of Liver Disorder Diseases by using Data Mining Technique: Survey"," International Journal of Pure and Applied Mathematics, 2018.
[12] A. S. S. M. I. ,. A. C. Thirunavukkarasu K., ""“Prediction of Liver Disease using Classification Algorithms “," 4th International Conference on Computing Communication and Automation (ICCCA), 2018.
[13] A. A. ,. W. A. A. D. S. Kanza Hamid, ""“Machine Learning with Abstention for Automated Liver Disease Diagnosis”," International Conference on Frontiers of Information Technology, 2017.
[14] O. S. G. A. K. M. A. Heba Ayeldeen, ""“Prediction of Liver Fibrosis stages by Machine Learning model ": A Decision Tree Approach," Third World Conference on Complex Systems (WCCS), 2015.
[15] V. G. S. D. H. Deepa H Belavigi, ""“Prediction of Liver Disease using Rprop, SAG and CNN”," International Journal of Innovative Technology and Exploring Engineering (IJITEE), pp. .ISSN: 2278-3075, Volume-8 Issue-8, June, 2019.
[16] S. K. Sanjay Kumar, ""“Effective Analysis and Diagnosis of Liver Disorder by Data Mining”," Proceedings of the International Conference on Inventive Research in Computing Applications, 2018.
[17] G. E. Somaya Hashem, ""“Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients“," IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. (Volume: 15 , Issue: 3, 2018.
[18] L. Z. S. C. S. A. Varun Vats, ""“A Comparative Analysis of Unsupervised Machine Techniques for Liver Disease Prediction”," IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 2018.
[19] D. D. Panigrahi Srikanth, ""“ A Critical Study of Classification Algorithms Using Diabetes Diagnosis”," in IEEE 6th International Conference on Advanced Computing, 2016.
[20] A. S. S. M. I. A. C. Thirunavukkarasu K., ""“Prediction of Liver Disease using Classification Algorithms”," 4th , p. International Conference on Computing Communication and Automation (ICCCA), 2018.
[21] M. Data, "Baiju Department of Information Technology Hindustan Institute of Technology and Science," [Online].
[22] G. C. U. M. L. R. Dua D, CA: University of California, School of Information and Computer Science, p. [http://archive.ics.uci.edu/ml, 2019.
[23] J. O. M. R. M. D. A. W. M.J. Sorich, "Comparison of Linear and Nonlinear Classification Algorithms for the Prediction of Drug and Chemical Metabolism by Human UDP-Glucurono syltransferase Isoforms," Journal of Chemical Information and Computer Sciences , p. pp2019:2024, 2003.
[24] F. Markowetz, "Klassifikationmit support vector Machines," http://lectures.molgen.mpg.de/statistik03/docs /Kapitel 16.pdf, 2003.

[25] E. Fix and J. L. Hodges, "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties (PDF) (Report)," USAF School of Aviation Medicine, Randolph Field, Texas, 1951.
[26] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression"," The American Statistician, p. 46 (3): 175–185, 1992.
[27] A. ". M. McCallum, "Bayesian Network Represention Graphical Models,," 2019.
[28] S. M. Piryonesi and T. E. El-Diraby, "Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems," Journal of Transportation Engineering, Part B: Pavements, 2020.
[29] D. J. Hand and K. Yu, International Statistical Review, 2001.
[30] Vijayan, Veena, and Aswathy Ravikumar. "Study of data mining algorithms for prediction and diagnosis of diabetes mellitus." International journal of computer applications 95, no. 17 (2014).
[31] Fatima, M., Srivastav, S., & Mondal, A. C. (2017). Prenatal stress and depression associated neuronal development in neonates. International Journal of Developmental Neuroscience, 60, 1-7.
[32] Smith, J. D., & Liu, A. Y. C. (1988). The induction, desensitization and de-induction of tyrosine aminotransferase by 8-bromo-cyclic AMP in rat hepatoma cells. Biochemical Journal, 251(1), 261-267.
[33] Quinlan, J. R. (2014). C4. 5: programs for machine learning. Elsevier.
[34] Nai-arun, Nongyao, and Rungruttikarn Moungmai. "Comparison of classifiers for the risk of diabetes prediction." Procedia Computer Science 69 (2015): 132-142.
[35] Soltani, Zahed, and Ahmad Jafarian. "A new artificial neural networks approach for diagnosing diabetes disease type II." Int J Adv Comput Sci Appl 7, no. 6 (2016): 89-94.
[36] Sahan, Seral, et al. "Applications of Artificial Immune Systems-The Medical Applications of Attribute Weighted Artificial Immune System (AWAIS): Diagnosis of Heart and Diabetes Diseases." Lecture Notes in Computer Science 3627 (2005): 456-468.
[37] Anand, A., & Shakti, D. (2015, September). Prediction of diabetes based on personal lifestyle indicators. In 2015 1st International Conference on Next Generation Computing Technologies (NGCT) (pp. 673-676). IEEE.
[38] Rani, A. S., & Jyothi, S. (2016, March). Performance analysis of classification algorithms under different datasets. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 1584-1589). IEEE.
[39] Li, L. (2014, November). Diagnosis of diabetes using a weight-adjusted voting approach. In 2014 IEEE International Conference on Bioinformatics and Bioengineering (pp. 320-324). IEEE.
[40] Bashir, S., Qamar, U., Khan, F. H., & Javed, M. Y. (2014, December). An efficient rule-based classification of Diabetes using ID3, C4. 5, & CART ensembles. In 2014 12th International Conference on Frontiers of Information Technology (pp. 226-231). IEEE.Bas