International Journal of Advanced Research and Review

www.ijarr.org

Hepatitis Disease Prognosis Using Random Forest, K-Nearest Neighbors, Naive Bayes, Support Vector Machine, and Multi-Layer Perceptron

[1] M. Nageshwarappa, [2] Swetha,

[1] Assistant Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.
[2] MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

Abstract—

Among the leading infectious killers globally right now is hepatitis. In humans, it causes inflammation of the liver. We have a great opportunity to save many lives by detecting this fatal condition early on. In this study, we used several data mining approaches to forecast the occurrence of hepatitis. In addition to this, we have put out a respectable method for enhancing the accuracy of our prediction models. We eliminated observations with missing values as a means of dealing with missing data in our dataset. Using the info-gain feature selection technique in conjunction with ranker search, we were able to identify the characteristics that were not needed. The hepatitis illness dataset is used to determine the prediction accuracy using classification approaches such as K-Nearest Neighbors (KNN), Naive Bayes Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and Random Forest. To assess the classification models' performance, we have examined factors such as accuracy, precision, recall, F1-score, and ROC. Our prediction models are now more accurate thanks to the removal of missing-value data and the info-gain feature selection approach. Random Forest demonstrated the highest level of performance, with a classification accuracy of 92.41%. Relevant Search Terms—Hepatitis Disease, Data Mining, KNN, Naive Bayes, SVM, MLP, Random Forest

I. INTRODUCTION

The liver is a crucial organ in humans because of the numerous critical jobs it performs. Liver damage occurs as a result of hepatitis disease. Death may occur as a result of this cause. Early diagnosis of hepatitis illness in patients is a difficult problem for medical professionals. Currently, if we look at the healthcare sector, we can see that the quantity of data pertaining to health is growing daily. Data mining is a subfield of machine learning that helps researchers make informed decisions by effectively managing and solving complicated problems using large datasets. When used to large datasets, it may reveal previously unseen patterns and extract useful information. This is why it is the go-to method for researchers looking to address real-world issues. Nonetheless, many clinical reports and diagnostic test findings provide key information for the health care industry. A class name may be retrieved from a dataset using this method, which involves seeing an undetected pattern and paying 31.00 © in 2021. associated characteristics found in the dataset as per IEEE standards. The presence or absence of hepatitis illness in a patient may be determined using either the concealed pattern or the linked characteristics. It operates in a way that is analogous to an expert system. In addition to this, it will reduce the time and money needed for diagnosis. But numerous machine learning techniques are used for forecasting. Determining the optimal method is a challenging task for us. Our study used a variety of machine learning algorithms to determine if a patient has hepatitis. These algorithms include KNN, Naive Bayes, Support Vector Machine, Multi-Layer Perceptron, and

Random Forest. Our research makes three important contributions. As a preliminary step, we culled 155 hepatitis illness diagnostic datasets from the UCI machine learning collection, each with its own unique set of characteristics. Second, we improved the performance of our classification model by eliminating superfluous features using the info-gain feature selection process in conjunction with ranker search. Thirdly, we have compared the performance of our five methods, compared the results to those of prior research, and assessed the prediction outcomes according to various risk variables. Various portions have been included into our paper. Literature evaluation has been the main emphasis of Section II. We have covered approach in Section III. Experiment findings have been presented in Section IV. The last portion of our article has concentrated on its conclusion.

II. REVIEW OF LITERATURE

Using various data mining techniques, several studies have been conducted on the identification of hepatitis and other dangerous disorders. Using several decision tree methods, Manickam Ramasamy et al. [1] were able to forecast the occurrence of hepatitis illness. When compared to other methods, Random Forest has the highest accuracy at 87.50%. Using SVM and the Wrapper approach, A.H. Roslina et al. [2] were able to forecast the occurrence of hepatitis illness. With a gamma parameter value of 0.18, the best accuracy is 74.55%. The data mining tool Weka was used by the writers of this article for their study. "G. Sathya Devi" [3] identified cases of hepatitis using the CART, ID3, and C4.5 algorithms. The CART algorithm outperformed the competition with an accuracy rate of 83.2%. The data mining tool Weka was used by the writers of this article for their study. Logistic Regression, Decision Trees, Linear Support Vector Machine, and Naive Bayes Classifier were used by K. Santosh Bhargav et al. [4] to forecast the occurrence of hepatitis illness. Among the methods tested, Logistic Regression had the highest accuracy rate of 87.17 percent. In order to forecast progressive liver fibrosis in individuals with chronic hepatitis C, Somaya Hashem et al. [5] evaluated several methods. The ADT model has the highest accuracy of all of the prediction models tested so far at 84.4%. S. M. M. Hasan et al. [6] evaluated many supervised ML classification methods for the purpose of predicting the occurrence of cardiovascular disease. The info-gain feature selection strategy is used to increase the accuracy of the classification model in this article. Logistic Regression has shown to be the most accurate method, with a 92.76% success rate. The use of Support Vector Machines and the Chy-Square feature selection approach was used to forecast the occurrence of hepatitis illness by Varun Kumar et al. [7]. Support Vector Machines trained with the Chy-Square feature selection method achieved an accuracy of 83.12%. Using several decision tree methods, Nazmun Nahar and Ferdous Ara [8] were able to forecast the occurrence of liver illness. Decision Stump had the highest accuracy rate of all the methods tested (70.67%). Multiple categorization techniques were used by Md. Faisal Faruque et al. [9] to forecast Mellitus diabetes. The C4.5 decision tree outperformed all of the other methods with an accuracy of 73.5%. Using several classification methods, Tahira Islam Trishna et al. [10] identify hepatitis A, B, C, and E. Random Forest has produced the highest accuracy (98.6%) of all methods. In their study, Vedha Krishna Yarasuri et al. [11] used several classification algorithms to forecast the occurrence of hepatitis illness. The dataset is split into three parts: training set (60%), testing set (20%), and validation set (20%), according to this research. The highest accuracy, 96%, was achieved by ANN, outperforming all other methods. In their study, A. K. M. Sazzadur Rahman et al. [12] used several classification methods to forecast the occurrence of liver disease. With an accuracy of 74%, random forest outperforms all of the other methods.

III. METHODOLOGIES

Introduction to K-Nearest Neighbors (KNN) There are three stages to the classification process in this classifier. It calculates the K-value in step-1. Step two involves sorting the training data and computing the distance between each test sample; step three involves using a majority vote technique to assign a class name to the test sample data [6]. To get the distance in geometric terms, one uses:

$$E_d = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \qquad (1)$$

B. Naive Bayes

When it comes to classification, this method shines because to Bayes theorem. There is no learning curve for this classification model. Theoretically, we may use the probability of an event that has already happened to determine the likelihood of another event happening. To calculate the posterior probability, one must:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \qquad (2)$$

C. Support Vector Machine (SVM)

Classification and regression are two common uses for this supervised ML technique. For the objective of categorization, we have used this method in our publication. Pattern recognition from training data is its first step. Once a maximum margin hyperplane has been located, the data points are appropriately divided into their respective classes. We may use this hyperplane to make predictions based on the test data [13]. D. MLP–Multi-Layer Classification and regression are two common uses for this supervised ML technique. For the objective of categorization, we have used this method in our publication. The backpropagation approach is used to train the dataset first. Backpropagation is used to compute the gradient descent function that is employed to train the model. The model is able to anticipate the new test sample's class name after the training phase is over. There are three levels in an MLP network. The input layer is the first to take data. We move on to the concealed layer next. The number of them might be one or more. As a last step, the output layer produces the classified results [14]. Subject: Random Forest Classification and regression are two common uses for this supervised ML technique. We have classified using it in our study. There are three stages to do it. Several trees are used to create a forest of Decision Trees in the first stage of the learning process. For every set of test data, the trees that were used to create the forest in the previous phase will now predict a class name in step 2. Finally, in step 3, the test data is given the proper class name by a majority vote. Step 3 is encountered by all of the dataset's data points [6]. F. Method of Operation The following are the essential procedures for carrying out our research: First, create a file called dataset_1 and extract the dataset from the UCI machine learning repository. A comma-separated value file is what you're looking for. In Step 2, you will use RStudio to import the dataset 1.csv file. Thirdly, using the R programming language and the RStudio tools, verify the missing values. Step 4: Create a new file named dataset_2 and remove any entries that have an attribute with an empty value. Remember to save the file as a.csv file. There are no missing values in this file for any of the characteristics. Step 5: Next, open the dataset_2.csv file in Weka. Then, use the info gain feature selection technique and Ranker Search to identify highly correlated features. After that, create a new file called dataset_3 that only contains these features. The file extension is.csv as well. Step 6: Next, in order to determine whether hepatitis disease is present, load the dataset_1.csv file with all characteristics and missing values using the Weka classification tools. Then, use our five classifiers to categorize the dataset. Step 7: Next, use the Weka classification tools to import the dataset 2.csv file, which should have all characteristics and not have any missing values. Then, sort the dataset using our five classifiers to determine whether hepatitis disease is present or not. Step 8: Next, use the Weka classification tools to import the dataset 3.csv file. This file should only include strongly correlated features and should not have any observations with missing values. Then, classify the dataset using our five classifiers to determine whether hepatitis illness is present or not. Finally, in step 9, compare the results of the classification model that was obtained in steps 6–7 using dataset_2.csv, in step 8 using dataset_3.csv, and in step 6–7 using dataset_1.csv. In Step 10, we also evaluate the accuracy of our classification model by comparing it to the results of earlier studies.

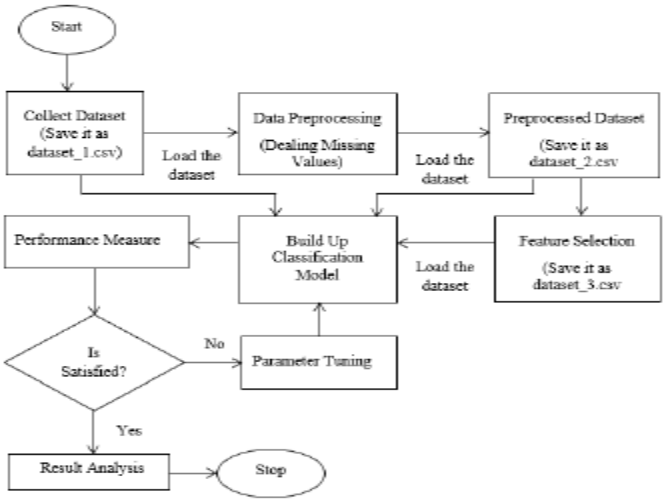

Fig 1. Flow chart of working procedure

IV. EXPERIMENTATION

We retrieved the dataset from the UCI ML repository for our study [15]. We have 155 entries in our dataset with a total of 20 characteristics. There are three distinct phases to our investigation. A. Preparing the Data In all, our dataset has twenty characteristics. These include factors such as age, sex, steroids, antivirals, lethargy, achy, anorexia, spiders, ascites, varices, bilirubin, alk phosphate, sgot, albumin, protime, histology, and class. There are a lot of blanks in our dataset. When dealing with missing values, there are various methods available. In order to train the classifier, we employed 79 data after first checking for records with missing values and removing such cases, following a prior study on hepatitis [1]. Reason being, missing value records may sometimes hurt categorization accuracy, thus we fixed them. Classification accuracy might be negatively impacted by missing variables. Our working process portion of the methodology section already discusses how we have dealt with records containing missing information in steps 3 and 4. On top of that, by using the info-gain attribute selection approach and Ranker Search, we can identify the dataset's strongly associated characteristics. The accuracy of the prediction model might be compromised if there are features in the dataset that are not strongly connected. Feature selection strategies have been used in our study for this same reason. The following 16 out of 20 attributes were chosen using this info gain feature selection method with Ranker Search: sex, steroids, antivirals, fatigue, malaise, anorexia, big liver, firm liver, spleen palpable, spiders, ascites, varices, albumin, protime, histology, and class. Step 5 of our work approach in the methodology section explains how we discovered strongly associated characteristics. B. Issue Description Since our study aims are comparable to those of Manickam Ramamasamy et al. [1], A. H. Roslina et al. [2], and Varun Kumar et al. [7], we have used their work as a foundation here. All twenty characteristics were used by Manickam Ramasamy et al. [1] for hepatitis prediction. Researchers Varun Kumar et al. [7] and A. H. Roslina et al. [2] have used feature selection techniques to identify strongly associated aspects. These features have been utilized to improve classification performance and make predictions about hepatitis illness. Ch. Findings and Analysis By using the Weka classification tools, five supervised classification algorithms were put into action. We have used 10-fold cross-validation to evaluate the performance of the model. Our categorization models have been applied to three criteria. Our CSV file, dataset_1, is first utilized with all 20 characteristics that have missing values. Next, we took all 20 characteristics from dataset_2 (our CSV file) that had complete values, and we picked out 16 attributes from dataset_3 (another CSV file). In steps four and five of the methodology section, we detailed our work technique and how we obtained the three CSV files that make up dataset_1, dataset_2, and dataset_3. Our study indicates that K=3 is the optimal choice for K-Nearest Neighbors. Our study indicates that the optimal value for Random Forest is numTrees=100. The optimal value for the parameter gamma in our study was 0.18, and we employed the Radial Basis Function (RBF) kernel for Support Vector Machines (SVM). One hidden layer has been used for Multi-Layer Perceptron. This hidden layer value is optimal. Additionally, the ideal values for momentum (0.2), training duration (400), and learning rate (0.01) are as follows. Table I, Table II, and Table III show the results for our five classifiers using our three datasets, respectively.

TABLE I. The result of Classification models using 20 Features Having Observations with Missing Values (dataset_1.csv)

Name of Classification Algorithm	Confusion Matrix		Accuracy
KNN	TP = 113	FN = 10	81.29%
	FP = 19	TN = 13	
Naive Bayes	TP = 110	FN = 13	85.16%
	FP = 10	TN = 22	
Support Vector Machine	TP = 123	FN = 0	79.35%
	FP = 32	TN = 0	
Multy-Layer Perceptron	TP = 112	FN = 11	84.52%
	FP = 13	TN = 19	
Random Forest	TP = 118	FN = 5	85.16%
	FP = 18	TN = 14	

TABLE II. The result of Classification models using 20 Features Having Observations without Missing Values (dataset_2.csv)

Name of Classification Algorithm	Confusion Matrix		Accuracy
KNN	TP = 61	FN = 5	82.28%
	FP = 9	TN = 4	
Naive Bayes	TP = 61	FN = 5	87.34%
	FP = 5	TN = 8	
Support Vector Machine	TP = 66	FN = 0	83.54%
	FP = 13	TN = 0	
Multy-Layer Perceptron	TP = 63	FN = 3	86.08%
	FP = 8	TN = 5	
Random Forest	TP = 65	FN = 1	89.87%
	FP = 7	TN = 6	

TABLE III. The result of Classification models using 16 Attributes Having Observations without Missing Values (dataset_3.csv)

Name of Classification Algorithm	Confusion Matrix		Accuracy
KNN	TP = 60	FN = 6	84.81%
	FP = 6	TN = 7	
Naive Bayes	TP = 61	FN = 5	88.61%
	FP= 4	TN = 9	
Support Vector Machine	TP = 66	FN = 0	91.14%
	FP = 7	TN = 6	
Multy-Layer Perceptron	TP = 62	FN = 4	84.81%
	FP = 8	TN = 5	
Random Forest	TP = 65	FN = 1	92.41%
	FP= 5	TN= 8	

Our performance comparison tables (I, II, and III) show that using 16 characteristics instead of 20 we were able to get better results from our classification models. Improving the efficiency of our categorization models is one of the benefits of removing records with blank values. In all three instances, the Random Forest approach outperforms the other four classification algorithms when comparing five different models. Figure 1 is a graphical representation of the accuracy of our categorization models:

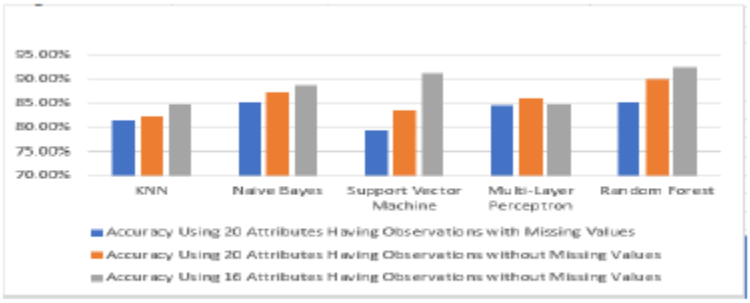

Fig. 2. Bar chart of the accuracy for our classifiers

TABLE IV. THE PERFORMANCE REPORT OF OUR CLASSIFIERS

Name of the Classification Algorithm	Precision	Recall	F1-Score	ROC Area
KNN (using 20 features having observations with missing values)	0.796	0.813	0.8	0.76
KNN (using 20 features having observations without missing values)	0.8	0.8	0.8	0.8
KNN (using 16 features having observations without missing values)	0.9	0.9	0.9	0.8
Naive Bayes (using 20 features having observations with missing values)	0.857	0.852	0.854	0.859
Naive Bayes (using 20 features having observations without missing values)	0.8	0.8	0.8	0.9
Naive Bayes (using 16 features having observations without missing values)	0.89	0.88	0.88	0.9
Support Vector Machine (using 20 features having observations without missing values)	0.63	0.79	0.70	0.2
Support Vector Machine (using 20 features having observations without missing values)	0.70	0.84	0.76	0.5
Support Vector Machine (using 16 features having observations without missing values)	0.92	0.91	0.897	0.73
Multi-Layer Perceptron (using 20 features having observations without missing values)	0.842	0.845	0.843	0.83
Multi-Layer Perceptron (using 20 features having observations without missing values)	0.84	0.86	0.85	0.85
Multi-Layer Perceptron (using 16 features having observations without missing values)	0.83	0.85	0.84	0.85
Random Forest (using 20 features having observations with missing values)	0.841	0.852	0.836	0.864
Random Forest (using 20 features having observations without missing values)	0.90	0.90	0.89	0.9
Random Forest (using 16 features having observations without missing values)	0.92	0.92	0.92	0.9

Compared to the 20 characteristics in datasets 1 and 2, Table-IV shows that the 16 attributes in dataset 3 performed better across the board in our classification models. It is not feasible to make accurate hepatitis predictions using all of the characteristics in our dataset. Because of this, we have identified and eliminated irrelevant features by using the info-gain attribute selection approach in conjunction with ranker search. Our classification models' performance has been enhanced in this manner. In a comparison of five different categorization models, the Random Forest method outperforms the other four with an accuracy of 92.41%. Beyond this, in the issue description part IV, we stated comparing the performance with past study. Our Random Forest model had a better accuracy of 92.41% than

that of Manickam Ramasamy et al. [1], whose best Random Forest model had an accuracy of 87.50%. The greatest accuracy obtained by A. H. Roslina et al. [2] using Support Vector Machine and the Wrapper approach was 74.55%; however, we outperformed them with 91.14% and 92.41% accuracy, respectively, using Random Forest and the info-gain feature selection strategy. Our best Support Vector Machine and Random Forest results are 91.14% and 92.41%, respectively, which is higher than the best Support Vector Machine and Chi-Square attribute evaluation results obtained by V. Kumar. M et. al. [7] (83.12%).

V. CONCLUSION AND FUTURE WORK

Our findings highlight the significance of feature selection and managing missing data for improving classification model accuracy. In order to find the most effective classifier, we compared all of our models. Using the info gain feature selection strategy with ranker search on our dataset and deleting observations from the dataset with missing values, each of our classification algorithms achieves exceptional performance. Low classification accuracy might be caused by missing values or features with little contribution in the dataset. Random Forest outperformed all five of our classifiers with a performance level of 92.41%. In order to carry out the experiment described in our research, we have used a tiny dataset. To further compare our categorization models in the future, it is recommended that we employ datasets of larger sizes. Additionally, alternative classification algorithms that use more effective methods of feature selection should be considered.

REFERENCES

[1] M. Ramasamy, S. Selvaraj, and Dr. M. Mayilvaganan. "An empirical analysis of decision tree algorithms: Modeling hepatitis data." IEEEInternational Conference on Engineering and Technology(ICETECH), India, pp. 1-4, 20 March. 2015.
[2] A. H. Roslina, and A. Noraziah. "Prediction of hepatitis prognosis using Support Vector Machines and Wrapper Method." IEEE Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), China, Vol. 5, pp. 2209-2211, 10-12 August. 2010.
[3] G. Sathyadevi"Application of CART algorithm in hepatitis disease diagnosis." IEEE International Conference on Recent Trends in Information Technology (ICRTIT), India, pp. 1283-1287, 3-5 June. 2011.
[4] K. S. Bhargav, T. D. Kumari, D. S. S. B. Thota, and V. B "Application of Machine Learning Classification Algorithms on Hepatitis Dataset." *International Journal of Applied Engineering Research*, vol. 13, no. 16, pp. 12732-12737, 2018.
[5] S. Hashem, G. Esmat, W. Elakel, S. Habashy, S. A. Raouf, M. Elhefnawi, M. El-Adawy, and M. Elhefnawi "Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients." *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 15 no. 3, pp. 861-868, 2018.
[6] S. M. M. Hasan, M. A. Mamun, M. P. Uddin, and M. A. Hossain, "Comparative Analysis of Classification Approaches for Heart Disease Prediction," IEEE International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, pp. 1-4, 8-9 February. 2018.
[7] V. Kumar. M, V. Sharathi. V, and G. D. BR. "Hepatitis prediction model based on data mining algorithm and optimal feature selection to improve predictive accuracy." *International Journal of Computer Applications*, vol. 51, no. 19, pp. 13-16, 2012.
[8] N. Nahar, and F. Ara. "Liver Disease Prediction by using Different Decision Tree Techniques." *International Journal of Data Mining & Knowledge Management Process (IJDKP),* vol. 8, no. 2, pp. 1-9, 2018.
[9] M. F. Faruque, Asaduzzaman, and I. H. Sarkar. "Performance analysis of Machine Learning Techniques to Predict diabetis Mellitus" IEEE International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox's Bazar, Bangladesh, pp. 1-4, 7-9 February. 2019.
[10] T. I. Trishna, S. U. Emon, R. R. Ema, G. I. H. Sajal, S. Kundu and T. Islam. "Detection of Hepatitis (A, B, C and E) Viruses Based on Random Forest, K-nearest and Naïve Bayes Classifier" IEEE International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, pp. 1-7, 6-8 July. 2019.
[11] V. K. Yarasuri, G. K. Indukuri and A. K. Nair. "Prediction of Hepatitis Disease Using Machine Learning Technique" IEEE International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, pp. 1-5, 12-14 December. 2019.
[12] A. K. M. S. Rahman, F. M. J. M. Shamrat, Z, Tasnim, J. Roy and S. A. Hossain. "A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms." *International Journal of Scientific & Technology Research (IJSTR),* vol. 8, no. 11, pp. 1-4, 2019.

[13] J. Cabrera, A. Dionisio and G. Solano. “Lung cancer classification tool using microarray data and support vector machines.” IEEE International Conference on Information, Intelligence, Systems and Applications (IISA), Corfu, Greece, pp. 1-6, 6-8 July. 2015.
[14] H. Yan, Y. Jiang, J. Zheng, C. Peng and Q. Li. “A multilayer perceptron-based medical decision support system for heart disease diagnosis.” *Expert Systems with Applications,* vol. 30, no. 2, pp. 272- 281, 2006.
[15] UCI Machine Learning Repository. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/hepatitis