Enhancing Liver Disease Diagnosis: Ensemble Techniques for Predictive Modeling and Diagnostic Strategies

Dr. Bhavesh M. Patel[1], Sirajbhai Abbasbhai Nagalpara[2]

[1] *Assistant Professor, Department of Computer Science, H.N.G.University, Patan, Gujarat, INDIA*
[2] *Scholar, Department of Computer Science, H.N.G.University, Patan, Gujarat, INDIA*

***Abstract*— Hepatitis C is a liver illness caused by a virus that is transferred by contact with contaminated blood, most commonly through drug use and needle sharing. It can result in chronic infection, leading to serious health issues like cirrhosis and liver cancer. Symptoms may not be apparent until the disease has progressed. Prevention involves avoiding behaviors that can spread the virus, and testing is essential as treatments can cure most cases within a few months. The objective is to leverage this data to predict liver disease in patients, providing significant benefits to both medical practitioners and individuals affected by the disease. Machine learning methods are employed due to the extensive amount of data available, enabling the utilization of past data to forecast future cases. To address this, the study proposes a performance optimization strategy that considers the training data and the variables that have a significant impact on the predictive model. Logistic Regression, SVM, KNN, Random Forest, Nave Bayes, and Stacking ensemble techniques are used to train the upgraded preprocessed data. Comparative analysis is conducted on the six models and against other research models. The novel model employing stacking classifier and surpasses others, achieving remarkable testing accuracy of 96%. This showcases our approach as a practical solution for real-world liver disease detection.**

***Keywords*— liver disease, Ensemble model, Data Mining, Classification Techniques**

I. INTRODUCTION

Hepatitis C is a result of infection with the hepatitis C virus (HCV) and is mainly spread through exposure to blood from an infected individual. Sharing needles or drug-related equipment is the main route of infection today. While some individuals experience a short-term illness, over 50% of those infected develop chronic hepatitis C, which can lead to severe health conditions like cirrhosis and liver cancer. Chronic hepatitis C often remains asymptomatic, and symptoms typically indicate advanced liver disease. Unfortunately, there is no available vaccine for hepatitis C. The key to prevention is avoiding behaviours that can spread the virus, particularly injection drug use. Testing for hepatitis C is crucial as treatments can cure the majority of cases within 8 to 12 weeks..[1]

Data mining extracts useful insights and information from large datasets using complex techniques developed from machine learning, statistics, and databases. By combining statistical and mathematical methodologies with artificial intelligence algorithms, it allows for the detection of important patterns and trends in data [1].

The basic purpose of data mining is to identify patterns that may be used to make sound business decisions. This technique holds significant importance in the advancement and expansion of companies, as it uncovers valuable insights and information that may be concealed within the data [2].

If not treated immediately, liver cancer, chronic kidney disease (CKD), breast cancer, cardiac syndrome, and diabetes offer serious health concerns and can be deadly. The healthcare industry may make educated and meaningful decisions by identifying underlying patterns and correlations in data [3].

[1] https:// www.cdc.gov/hepatitis/hcv/index.htm

II. BACKGROUND

A. Hepatitis C Virus

Persistent HCV infection results in liver cirrhosis and is linked to the occurrence of hepatocellular carcinoma (HCC). Each year, approximately 400,000 deaths are ascribed to HCV globally, with a significant impact observed in the United States [4].

In the absence of treatment, the majority of significant acute infections progress into chronic conditions, particularly in the case of liver-related diseases such as cirrhosis and liver cancer. Factors such as excessive alcohol consumption and metabolic syndrome play a significant influence in influencing the development and progression of liver diseases, including hepatocellular carcinoma (HCC) [5].

III. LITERATURE REVIEW

Günaydin at el. [6] according to the results, the best accuracy achieved was by using Decision Tree without image processing, with a rate of 93.24%. However, for cases where image processing was performed, Artificial Neural Networks provided the best accuracy rate of 82.43%.

Rao, G. at. el. [7] a classification experiment was carried out on the LIDC-IRDI dataset or prediction was made, and the corresponding accuracy was measured to be 84%.

Sharma, M. at. el. [8] the RFGBEL model has a high accuracy rate of 93.92%, a high sensitivity rate of 94.73%, an F-1 score of 0.93, a Log-Loss/Cross-entropy score of 5.89, and a Jaccard score of 0.72. The area under the curve is likewise computed by RFGBEL as 0.932.

Sagar Patel [9] it aims to categorize the stage of liver disease into four types: 1) Cirrhosis Liver, 2) Liver fibrosis, 3) Fatty Liver, and 4) Healthy Liver. In this context, the research compares several algorithms, including NB, SVM, LOR, RF, DT, KNN, and RBTC, with a newly suggested Hybrid Classifier (RF, SVC, XGBoost).

Sivasangari, A.. at. el. [10] approaches such as SVM (Support Vector Machines), RF (Random Forest), and DT (Decision Trees) are proposed to increase precision, accuracy, and reliability in predicting liver disease.

Wu, C. at. el. [11] research encompassed a sample of 577 patients, out of which 377 were diagnosed with fatty liver. The area under the receiver operating characteristic (AUROC) was calculated for each model using a 10-fold cross-validation technique. RF received 0.925, NB received 0.888, ANN received 0.895, and LR received 0.854. Furthermore, the corresponding accuracy rates were as follows: RF stood at 87.48%, NB at 82.65%, ANN at 81.85%, and LR at 76.96%.

Geetha, C. at. el [12] this study was to concentrate on classification algorithms used to distinguish healthy individuals within liver datasets. Additionally, based on their performance metrics, the research sought to compare these classification algorithms and furnish accuracy outcomes for prediction.

Atallah, R. at. el. [13] the patient is categorized using the collective decision of multiple machine learning models, aiming to enhance accuracy compared to relying on a single model. Ultimately, this method achieved a 90% accuracy by utilizing a hard voting ensemble model.

Anggraeni, M. at. el. [14] this study intends to improve the accuracy of chatbot answers by using an ensemble technique that combines five different machine learning categorization algorithms.

Ghaheri, P. at. el. [15] in this study, a level of accuracy reaching 85.42%, an F1-score amounting to 84.94%, precision at 86.77%, specificity of 87.62%, and sensitivity reaching 83.20% were attained. The research results highlighted that this approach surpassed existing methods and holds potential to aid medical professionals in Parkinson's disease diagnosis.

Su, H. at. el. [16] the validity of the suggested techniques is confirmed through the utilization of three actual hyperspectral datasets. The practical results show that TCRC-bagging and TCRC-boosting outperform their individual classifier equivalents.

Khamparia, A. at. el. [17] a combination of deep learning multi-model ensemble techniques was applied to two datasets concerning neuromuscular disorders. This approach led to enhanced predictive accuracy compared to different datasets and classification methods.

Gao, X. at. el. [18] a comparison is made between machine learning algorithms and ensemble learning techniques, focusing on specific characteristics. Various methods are utilized to assess the models, incorporating metrics such as accuracy, recall, precision, F-measure, and ROC curves. The results suggest that the combination of the bagging ensemble learning method and decision trees has shown the highest level of performance.

IV. RESEARCH METHODOLOGY

This section provides a summary of the datasets, the suggested method, the structural design of the system, and the algorithms utilized for categorizing liver disease.

A. Dataset Description

Liver disease categorization is performed using the dataset pertaining to Indian Liver Patients (ILPD) sourced from the UCI Machine Learning Repository.[2] It comprises 13 columns. Table 1 presents a summary of the feature characteristics for the patients.

B. Dataset Pre-Processing

The target feature in the dataset represents the categorical health condition of patients' livers. The control group, also known as the negative class, consists of blood donors. On the other hand, the positive classes include patients diagnosed with Cirrhosis, Fibrosis, or Hepatitis. Additionally, there is a smaller subset of patients categorized as suspect blood donors. The objective of the machine learning model is to predict the appropriate class to which a patient belongs, based on the given information. 0: Blood Donors, 0s: Suspect Blood Donors, 1: Hepatitis Hepatitis, 2: Fibrosis, 3: Cirrhosis. The dataset has 31 missing values, the majority of which are located in the 'ALP' and 'CHOL' columns.

C. Figures and Tables

a) Explanation of the table found in the HCV dataset.:

ID	Attribute	Description	Processing
1	Category	Categoric	0: Blood Donors 0s: suspected Blood Donors, 1: diagnosed Hepatitis 2: Fibrosis 3: Cirrhosis
2	Age	Numerical	Min:19 Max:77 Mean:47.41 Std:10.06
3	ALB	Numerical	Min:14.9 Max:82.2 Mean:41.62 Std:5.78
4	ALP	Numerical	Min:11.3 Max:416.6 Mean:68.28 Std:26.03
5	ALT	Numerical	Min:0.9 Max:325.3 Mean:28.45 Std:25.47
6	AST	Numerical	Min:10.6 Max:324.0 Mean:34.79 Std:33.09
7	BIL	Numerical	Min:0.8 Max:254.0 Mean:11.4 Std:19.67
8	CHE	Numerical	Min:1.42 Max:16.41 Mean:8.2 Std:2.21
9	CHOL	Numerical	Min:1.43 Max:9.67 Mean:5.37

[2] https://archive.ics.uci.edu/ml/datasets/HCV+data

			Std:1.13
10	CREA	Numerical	Min:8.0 Max:1079.1 Mean:81.29 Std:49.76
11	GGT	Numerical	Min:4.5 Max:650.9 Mean:39.53 Std:54.66
12	PROT	Numerical	Min:44.8 Max:90.0 Mean:72.04 Std:5.4
13	sex	Nominal	'f'=female, 'm'=male

D. Approach and Structure for liver diseases

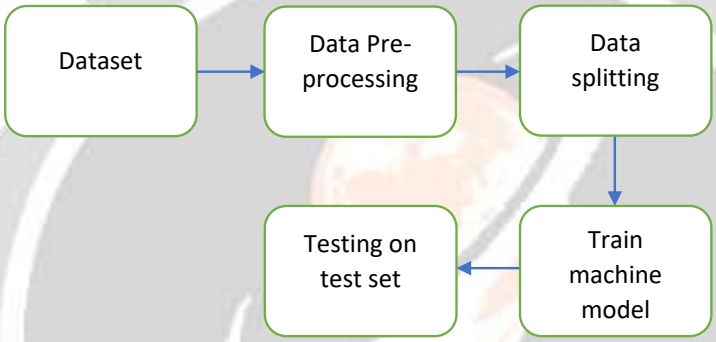


Figure 2: The suggested design framework.

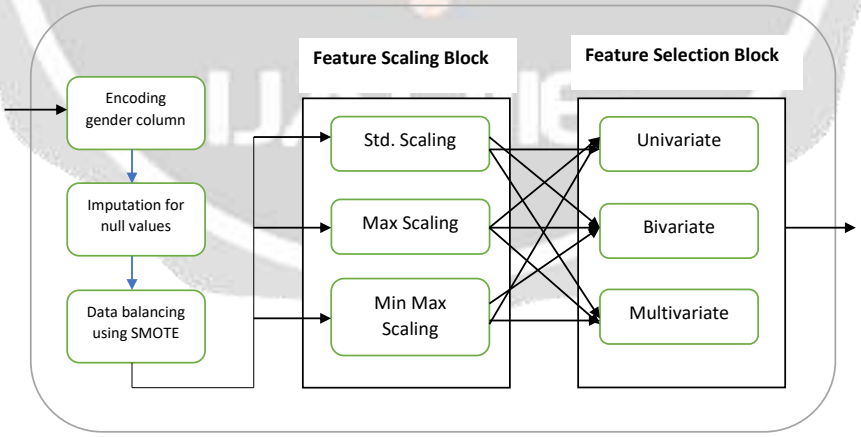


Figure 3: The suggested initial processing stage for liver disease classification.

Data Balancing using SMOTE

Before SMOTE

No.	Attribute	Ratio
1	Health	86.5%
2	Hepatitis	4.4%
3	Fibrosis	3.7%

4	Cirrhosis	5.3%

After SMOTE

No.	Attribute	Ratio
1	Health	25.0%
2	Hepatitis	25.0%
3	Fibrosis	25.0%
4	Cirrhosis	25.0%

E. Utilizing Improved Preprocessing for Liver Disease Prediction via Machine Learning Algorithms.

This study assesses how ensemble-driven machine learning techniques perform on the Dataset and conducts a comparative analysis of their outcomes. The Ensemble methodology involves a distinct strategy where we merge several machine learning models, whether similar or dissimilar, to execute prediction tasks, such as logistic regression (LR), KNN, support vector machines (SVM), and so forth [19]. The ensemble models employ foundational estimators or base learners. There exist numerous rationales for favouring ensemble models over conventional ones.

Our results indicate that the utilization of ensemble classification approaches leads to higher accuracy when compared to individual classifiers [20]. The amalgamation of these algorithms demonstrated superior performance in contrast to using a single algorithm. The discovery was made that selecting classifiers with independence and divergent perspectives leads to enhanced outcomes [21].

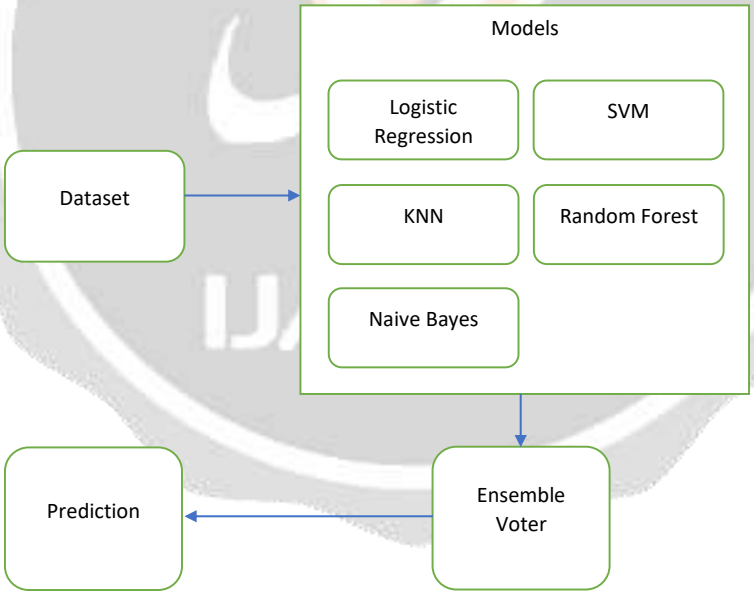


Figure 4: Ensemble learning involves training data using diverse base classifiers, and their outputs are merged to derive the ultimate prediction.

Table 2 outlines the process of fine-tuning hyperparameters for all ensemble learning models with OPTUNA.

Models	Hyperparameters	Optimal Value
Random Forest	n_estimators:['100, 150, 200, 500 '], split_min_samples:['1.0, 2, 4, 5'], criterion:['gini, entropy '] , leaf_node_max:['4, 10, 20, 50, None'] min_leaf_samples:['1, 2, 4, 5']	100 2 gini None 1
Logistic Regression	n_estimators:['100, 150, 200, 500'],	100

	leaf_min_samples:['1, 2, 4, 5'], criterion: ['l1, l2, elasticnet'], leaf_node_max: ['4, 10, 20, 50, None'], min_leaf_samples:['1.0, 2, 4, 5'],	1 elasticnet None 2
KNN	n_estimators:['100, 150, 200, 500'], criterion:['uniform, distance '], split_min_samples:['1.0, 2, 4, 5'], leaf_node_max:['4, 10, 20, 50, None'], min_samples_leaf: ['1, 2, 4, 5'],	100 distance 2 None 1
SVM	n_estimators:['100, 150, 200, 500'], criterion:['poly, rbf '] , leaf_min_samples:['1, 2, 4, 5'], split_min_samples:['1.0, 2, 4, 5'], leaf_node_max:['4, 10, 20, 50, None']	100 poly 1 2 None
Naïve Bayes	n_estimators:['100, 150, 200, 500'], criterion:['nb_smoothing']	100
Ensemble	n_estimators:['lr, knn, svm, rf, nb'], direction=['maximize '] , n_trials= ['100']	100

F. Experimental Setup

The research conducted in this study involved conducting experiments on a personal computer running the Windows 11 operating system. The computer used in the experiments has the following hardware specifications: 8 GB of RAM and a CPU from the Intel i5-10th generation. External GPUs were not utilized in this setup. The Python programming language was used in a Jupyter Notebook environment to carry out the coding. The research employed a variety of widely recognized machine learning libraries, including pandas, numpy, sklearn, and seaborn, to facilitate the execution of the tasks.

G. Metrics of Evaluation

Six performance criteria are used to test and validate the proposed model: accuracy, precision, recall, and F-Measure [22]. Eq. 2-5, corresponding formulas are used to compute these measurements [23].

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

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

$$F1 = \frac{2 \times Precision \times Recall}{Precision+Recall} \quad (4)$$

$$Accuracy = \frac{TP+TN}{Total\ Instances} \quad (5)$$

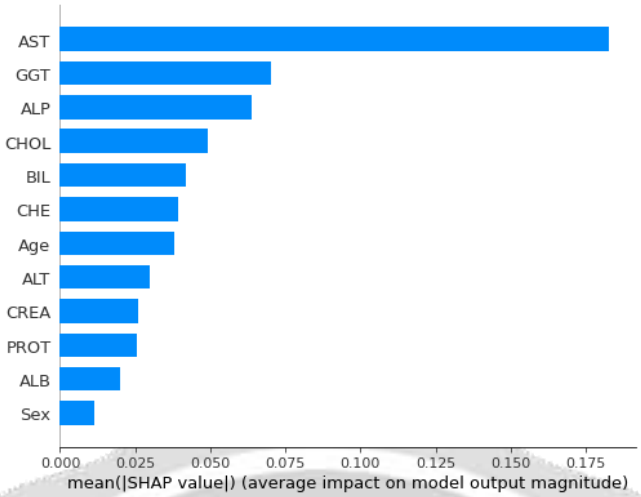
TP denotes true positives, FP denotes false positives, FN denotes false negatives, and TN denotes true negatives [22].

V. EXPERIMENTAL RESULTS

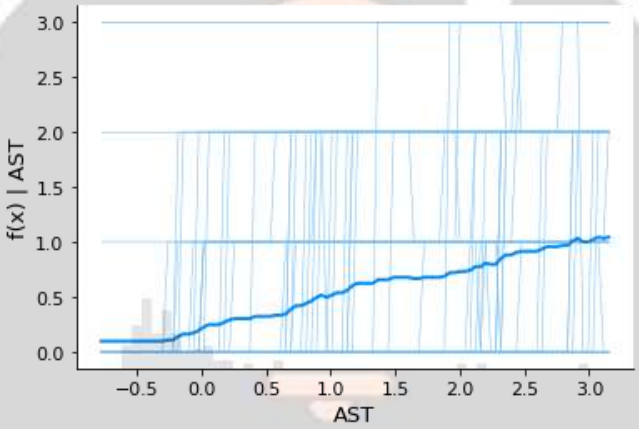
Table 3 shows the model's performance.

Measure	Accuracy	Precision	Recall	f-measure
Ensemble	96	98	100	99
Logistic Regression	88	100	89	94
KNN	90	99	92	95
SVM	91	97	94	96
Random Forest	83	97	86	91

Naïve Bayes	83	97	86	91



It calculates and visualizes how changes in the chosen feature affect the model's predictions, while keeping other features constant. This aids comprehension of the link between the chosen feature and the model's output.



VI. DISCUSSION AND CONCLUSION

This research centers on creating a predictive model based on ensemble techniques, employing multiple classifiers to forecast and categorize liver disease. The researchers use the chronic renal illness dataset from UCI's machine learning repository to do this, as well as pre-processing methods to cope with any missing data.

The results reveal that the ensemble-based prediction model beats the current models, with a 96% accuracy. As a result, the model is regarded as an effective predictor of liver disease.

REFERENCES

[1] The paper was written by A. Al-Aiad, S. Abualrub, Y. Alnsour, and M. Alsharo and was titled "Data Mining Algorithms Predicting Different Types of Cancer: Integrative Literature Review." It debuted during the AMCIS 2020 TREOs. You may find the document at https://aisel.aisnet.org/treos_amcis2020/59.

[2] R. D. Canlas Jr. finished an unpublished master's thesis titled "DATA MINING IN HEALTHCARE: CURRENT APPLICATIONS AND ISSUES" in August 2009. The ten-page thesis focuses on the use of data mining in healthcare.

[3] Ibrahim and A. Abdulazeez wrote a paper in the Journal of Applied Science and Technology Trends titled "The Role of Machine Learning Algorithms in Disease Diagnosis." The essay appears on pages 10 through 19 of volume 2, issue 1. It was published in 2021 and has the following DOI: 10.38094/jastt20179.

[4] "Hepatitis C Virus Vaccine: Challenges and Prospects," co-authored by J. D. Duncan, R. A. Urbanowicz, A. W. Tarr, and J. K. Ball, was published in the journal "Vaccines." The paper goes from page 1 through page 23 of volume 8, issue 1. It was published in 2020 and has the DOI: 10.3390/vaccines8010090.

[5] L. Syafa'ah, Z. Zulfatman, I. Pakaya, and M. Lestandy did study titled "Comparison of Machine Learning Classification Methods in Hepatitis C Virus," which was published in 2021 in the Journal of Online Information, volume 6, issue 1, page 73. The corresponding DOI is 10.15575/join.v6i1.719.

[6] Günaydin, M. Günay, and engel co-authored a paper titled "Comparison of Lung Cancer Detection Algorithms," which was presented at the 2019 Scientific Meeting on Electrical, Biomedical Engineering, and Computer Science. The DOI for the publication is 10.1109/EBBT.2019.8741826, and it is tied to the EBBT 2019 event.

[7] G. S. Rao, G. V. Kumari, and B. P. Rao contributed to "Network for Biomedical Applications," which was published by Springer Singapore in Volume 2, Issue 1 in January 2019. The DOI for this article is 10.1007/978-981-13-1595-4.

[8] "Enhanced Prognosis of Hepatocellular Carcinoma Fatality Using Ensemble Learning Approach," by M. Sharma and N. Kumar, was published in the Journal of Ambient Intelligence and Humanised Computing. The paper was published in Volume 13, Issue 12, pages 5763-5777 in 2022. The associated DOI is 10.1007/s12652-021-03256-z.

[9] Mr. Sagar Patel of D. P. P. and Dr. Chintan Shah of Dr. Chintan Shah did study on "Diagnosis of Liver Diseases and Prediction of Liver Disease Stage Using Hybrid Machine Learning Classifiers." This research appears in Volume 38, Issue 3, pages 945-954, published in 2023. 10.5281/zenodo.7923033 is the DOI.

[10] M. Banu Priya, P. Laura Juliet, and P. R. Tamilselvi conducted research on the "Evaluation of Liver Disease Prediction Using Machine Learning Algorithms." The study was published in the International Research Journal of Engineering and Technology in 2018, Volume 5, Issue 1, pages 206-211. The story may be found at www.irjet.net.

[11] "Machine Learning Algorithms for Predicting Fatty Liver Disease," by C. C. Wu et al., was published in the journal "Computational Methods and Programmes in Biomedicine." The article was published in 2019 and can be found on pages 23-29 of Volume 170. The corresponding DOI is 10.1016/j.cmpb.2018.12.032.

[12] C. Geetha and A. R. Arunachalam presented their study "Approaches for Evaluating Liver Disease Prediction Using Machine Learning Algorithms" at the 2021 International Conference on Computer Communications and Informatics (ICCCI 2021). In 2021, the work was published in the conference proceedings on pages 55-58. The DOI for this work is 10.1109/ICCCI50826.2021.9402463.

[13] R. Atallah and A. Al-Mousa presented a work titled "Heart Disease Detection Using a Majority Voting Ensemble Method Based on Machine Learning" at the 2019 2nd International Conference on New Trends in Computer Science (ICTCS 2019). In 2019, the study was published on pages 1-6 of the conference proceedings. The DOI for this paper is 10.1109/ICTCS.2019.8923053.

[14] M. Anggraeni and H. A. Damanik authored a paper titled "Integrating Multiple Models for Accurate Answer Prediction in Chatbots using an Ensemble Method" in the "Journal of Research on Post and Informatics." The article is in Volume 11, Issue 2, spanning pages 137–152, published in 2021. The associated DOI is 10.17933/jppi.v11i2.352.

[15] P. Ghaheri, H. Nasiri, A. Shateri, and A. Homafar were among those who contributed to the study "Parkinson's Disease Diagnosis Based on Voice Signals Using SHAP and Hard Voting Ensemble Method." This study, which runs from pages 1 to 19, is accessible online at http://arxiv.org/abs/2210.01205 and will be published in 2022.

[16] H. Su, Y. Yu, Q. Du, and P. Du, "Ensemble Learning for Hyperspectral Image Classification Using Tangent Collaborative Representation," IEEE Trans. Geosci. Remote Sens., vol. 58, no. 6, pp. 3778-3790, 2020, doi: 10.1109/TGRS.2019.2957135.

[17] A. Khamparia, A. Singh, D. Anand, D. Gupta, and A. Khanna gave a presentation titled "Prediction of Neuromuscular Disorders Using a Novel Deep Learning-Based Multi-Model Ensemble Method." The manuscript may be found on page 11095 of the journal's volume 0 publication. It will be launched in 2020.

[18] X conducted a research project. A. A. Ali, H. S. Hassan, and E. Gao. M. Anwar with the purpose of "Enhancing Accuracy in Predicting Heart Diseases Through the Use of Ensemble Method." The research was published in volume 2021 in the year 2021.

[19] The paper "Improved Preprocessing Approach Using Ensemble Machine Learning Algorithms for Liver Disease Detection" was co-authored by A. Q. Md, S. Kulkarni, C. J. Joshua, T. Vaichole, S. Mohan, and C. Iwendi. In the year 2023, the work was published in the journal "Biomedicines," Volume 11, Issue 2. The DOI for this paper is 10.3390/biomedicines11020581.

[20] J. L. Fernandez-Aleman, J. M. Carrillo-De-Gea, M. Hosni, A. Idri, and G. Garcia-Mateos wrote "Ensemble Classification Methods in Diabetes Disease: A Homogeneous and Heterogeneous Analysis." In 2019, the IEEE Engineering in Medicine and Biology Society (EMBS) Annual International Conference included this review. This publication's page range is 3956-3959, and the DOI is

10.1109/EMBC.2019.8856341.

[21] M. Ahamad and N. Ahmad published "Utilisation of Ensemble Methods for Assessing Students' Knowledge" in the journal "International Journal of Information Technology." The paper was published in 2021 in Volume 13, Issue 3, pages 1025-1032. The corresponding DOI is 10.1007/s41870-020-00593-8.

[22] M. N. Hossin and M. Sulaiman reviewed "Evaluation Metrics for Data Classification Assessment" in October of 2010. The article may be found on pages 4-5 of volume 0, issue October.

[23] W. Wang and Y. Lu conducted research on "Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) in Evaluating Rounding Models." The paper was published in the IOP Conference Series: Materials Science and Engineering, volume 324, number 1, in 2018. The DOI is 10.1088/1757-899X/324/1/012049.