Machine Learning for Classifying Liver Diseases

Omar Ahmed
Department of Artificial Intillegence
Misr University for Science and Technology
Cairo, Egypt
98073@must.edu.eg

Rana Shiba
Department of Artificial Intelligence
Misr University for Science and Technology
Cairo, Egypt
89597@must.edu.eg

Philopateer Magdy
Department of Computer Science
Misr University for Science and Technology
Cairo, Egypt
94286@must.edu.eg

***Abstract*— Hepatitis C poses a significant public health challenge, often progressing to life-threatening chronic infections without early symptoms, making it difficult to detect and diagnose This study promptly provides a solution by developing a machine learning model aimed at the early detection of hepatitis C-related liver diseases A comprehensive analysis was done using data from the captured UCI Machine Learning Repository, with laboratory internal values and patient demographics and apply 5 algorithms for liver diseases classification. The study's results were good, achieving a validation accuracy of 97% and testing accuracy of 95%. For detailed analysis, we used a confusion matrix to reveal additional insights into the model performance, as it achieved a Precision score of 99%, Recall score of 99%, and F1-score of 99% as well. These findings present the promising potential of machine learning for early disease classification, which can help in medical applications, minimize human error, and participate in improved detection of Liver diseases.**

Keywords:Machine learning , Hepatitis C , UCI Dataset

I. INTRODUCTION

The Centers for Disease Control and Prevention (CDC) outlines a public health challenge posed by a liver infection, Hepatitis C, which is caused by Hepatitis C virus. Hepatitis C is transmitted through contact with an infected blood due to needle-sharing or drug-injecting practices. A significant proportion of the infected individuals can progress to a life-threatening chronic infection.

The chronic infection progresses without symptoms resulting in high complexities in early detection and diagnosis. When symptoms arise, they indicate an advanced liver disease which undermines the importance of proactive testing. Due to the unavailability of a vaccine, prevention of the virus is mainly highlighted by avoiding behaviors that result in blood contact with an infected person, particularly the injection of drugs, which transmits the disease.

Due to the discussed challenges, this study aims to propose a forward-looking solution: the creation of a predictive model for the early detection of Hepatitis C and other liver diseases. Such solution could enhance disease management by allowing individuals quickly assess their risk of the virus progression and seek immediate treatment. The subsequent sections discuss the dataset utilized, the methodologies involved in predictive modeling, and the potential impacts that the model could have on the public health and on the individual's well-being.

II. METHODOLOGY

The methodology employed in this code sequence encompasses a systematic approach to data analysis,

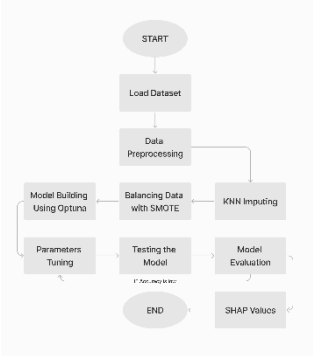

Fig.1 Flowchart of the System

preprocessing, modeling, and evaluation. The following steps outline the key processes involved in achieving a comprehensive understanding of the dataset and developing a robust machine learning model:

We started by loading the dataset and getting to know it, and it was revealed that the dataset was unbalanced as the normal class represented 86.5% of the dataset and the other three classes represented the rest, and it included some NULL values as well. So, we Used KNN imputing to fill those NULL values as it's a critical step that must be done before dealing with the data being unbalanced. Then we used SMOTE (Synthesized Minority Oversampling Technique) to make the dataset balanced and more qualified to be fed to our machine learning model.

After preprocessing the data, we moved on to the next step which was creating the model. By using Optuna, we were able to create an Ensemble Model that consists of five different algorithms: K-Nearest Neighbor (KNN), Logistic Regression (LR), Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF). The hyperparameters of each model were adjusted and tuned using Optuna, and then an ensemble voter was employed to extract the best prediction.

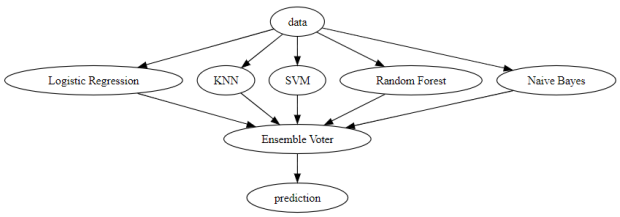

Fig.2 Ensemble Model

III. DATASET

The dataset utilized in this study originates from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/HCV+data.

It contains laboratory values of blood donors and Hepatitis C patients and demographic values like age. There are 615 records in the original dataset and 12 features aside from the target column. The columns are:

- Category: The target feature. values: '0=Blood Donor', '0s=suspect Blood Donor', '1=Hepatitis', '2=Fibrosis', '3=Cirrhosis'
- Age: age of the patient in years
- Sex: sex of the patient ('f'=female, 'm'=male)
- ALB: amount of albumin in patient's blood
- ALP: amount of alkaline phosphatase in patient's blood
- ALT: amount of alanine transaminase in patient's blood
- AST: amount of aspartate aminotransferase in patient's blood
- BIL: amount of bilirubin in patient's blood
- CHE: amount of cholinesterase in patient's blood
- CHOL: amount of cholesterol in patient's blood
- CREA: amount of creatine in patient's blood
- GGT: amount of gamma-glutamyl transferase in patient's blood
- PROT: amount of protein in patient's blood

	Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
1	0=Blood Donor	32	m	38.5	52.5	7.7	22.1	7.5	6.93	3.23	106.0	12.1	69.0
2	0=Blood Donor	32	m	38.5	70.3	18.0	24.7	3.9	11.17	4.80	74.0	15.6	76.5
3	0=Blood Donor	32	m	46.9	74.7	36.2	52.6	6.1	8.84	5.20	86.0	33.2	79.3
4	0=Blood Donor	32	m	43.2	52.0	30.6	22.6	18.9	7.33	4.74	80.0	33.8	75.7
5	0=Blood Donor	32	m	39.2	74.1	32.6	24.8	9.6	9.15	4.32	76.0	29.9	68.7
6	0=Blood Donor	32	m	41.6	43.3	18.5	19.7	12.3	9.92	6.05	111.0	91.0	74.0
7	0=Blood Donor	32	m	46.3	41.3	17.5	17.8	8.5	7.01	4.79	70.0	16.9	74.5
8	0=Blood Donor	32	m	42.2	41.9	35.8	31.1	16.1	5.82	4.60	109.0	21.5	67.1
9	0=Blood Donor	32	m	50.9	65.5	23.2	21.2	6.9	8.69	4.10	83.0	13.7	71.3
10	0=Blood Donor	32	m	42.4	86.3	20.3	20.0	35.2	5.46	4.45	81.0	15.9	69.9

Fig.3 Sample of the Dataset

The following figure shows a scatter plot of the dataset with all its classes present in the visualization:

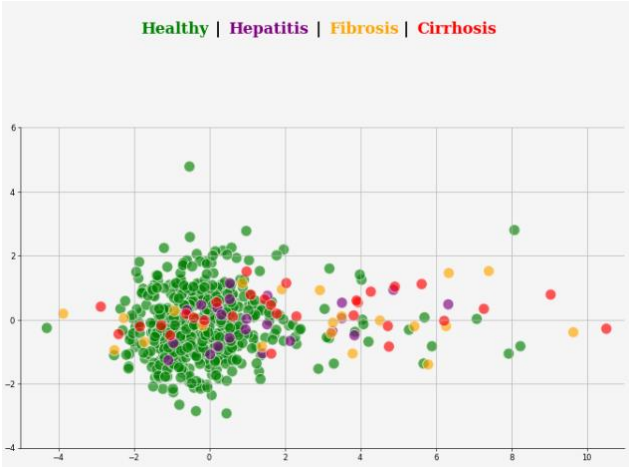

Fig.4 Dataset Scatter Plot

And the next plot shows each class in a separate plot:

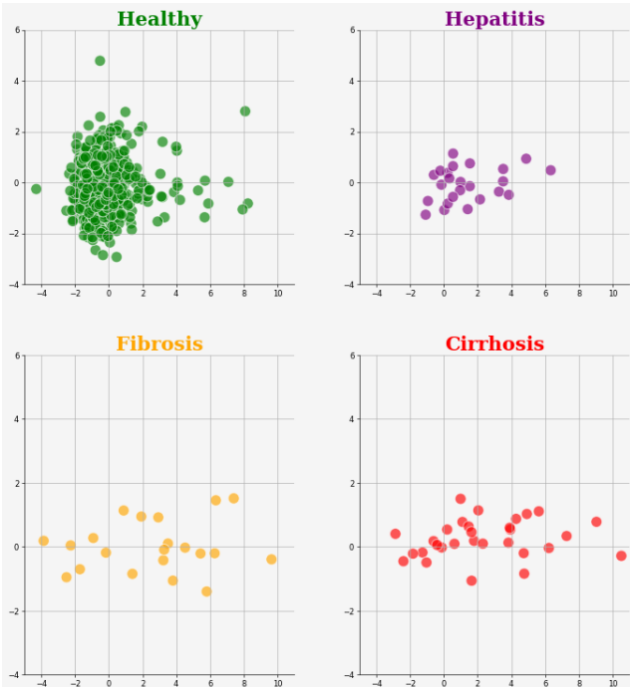

Fig.5 Scatter Plot of All Classes

As it's demonstrated from the above figure, the dataset is unbalanced and needs to be balanced, so after using SMOTE for balancing the dataset, we made all classes the same size as previewed in the following figure:

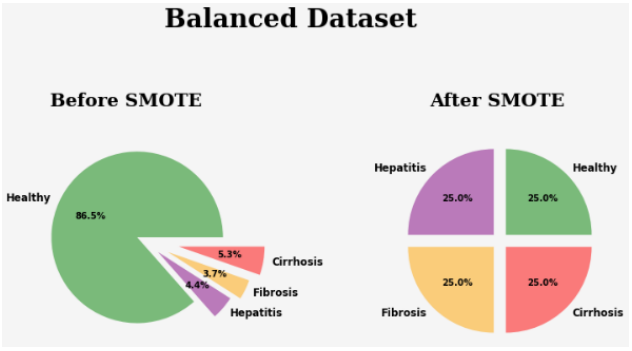

Fig.6 Dataset Balancing Using Smote

IV. RELATED WORKS

Machine learning algorithms can be implemented in various fields for enhancing the prediction of a targeted feature. Aiming to utilize machine learning in the medical field, we hence sought related works that detect liver diseases using different models based on imbalanced datasets.

Starting with a machine learning system by Mounita G. et al, the aim of the study is to evaluate the usage of different algorithms including logistic regression, random forest, XGBoost, support vector machine (SVM), AdaBoost, K-NN, and decision tree to predict the diagnosis of a chronic liver disease. The evaluation of the best algorithm approach was done through calculating accuracy, precision, recall, F1 score, an area under the curve (AUC), and specificity. Based on the latter, the random forest algorithm showed the best performance with 83.70% accuracy in liver disease prediction. Adding to high accuracy, random forest showed a better score in precision,

F1, recall, and AUC metrics compared to the other algorithms tested.

Another study done by Bilal K. et al aims to implement a new solution to predict early analysis of liver diseases based on Composite Hypercube on Iterated Random Projection (CHIRP). CHIRP is a classifier designed to address the curse of dimensionality and exponential complexity by using projection, binning, and covering in a sequential framework. Similar to the study by Mounita G., this system compares previously used models that are based on MLP, KNN, SVM, J48, RF, DS, RT and LR and evaluates their performance using accuracy, Mean Absolute Error (MAE), Relative Absolute Error (RAE). Based on the comparison results, CHIRP performs the best in reducing the error rate in assessment measurements rather than another utilized models.

Lastly, a study by S. Katiyar et al. proposes a system to assess the efficiency and accuracy of various models in predicting chronic liver diseases. Structured data was used in comparing machine learning classification models including XGBoost Classifier, random forest, decision tree, and logistic regression. Based on accuracy metrics, logistic regression and XGBoost Classifier showed the highest performance in prediction. Random forest resulted in accuracy about 74.57% which is impractical due to low accuracy.

V. Results & Discussion

In this study, our primary objective was to employ machine learning methods for the classification of liver diseases, namely Cirrhosis, Fibrosis, and Hepatitis-C.

To achieve the highest performance while maintaining computational demands, we adopted SMOTE for dataset balancing and Optuna for ensemble model implementation. We achieved validation accuracy of 97% and testing accuracy of 95%.

The best hyperparameters for each algorithm was as follows:

Table 1 Hyperparameters of Each Algorithm

Algorithm	Parameters
Logistic Regression	'lr_penalty': 'l2' 'lr_solver2': 'saga' 'lr_tol': 0.0064830531932500695 'lr_C': 0.8800756586994839 'lr_w': 0.12388677262046863
K-Nearest Neighbors	'knn_neighbors': 55 'knn_weights': 'distance' 'knn_p': 2 'knn_w': 0.5741122613065569
Support Vector Machine	'svm_C': 0.6902832081216106 'svm_kernel': 'poly' 'svm_degree': 4 'svm_tol': 0.00204169447396070 'svm_w': 0.3772048225616136
Random Forest	'rf_estimators': 64 'rf_criterion': 'entropy' 'rf_max_depth': 86 'rf_min_samples_split': 8 'rf_min_samples_leaf': 11 'rf_w': 0.8789688421090811
Naïve Bayes	'nb_smoothing': 5.8605852654821e-07 'nb_w': 0.12001454957720925

We then used the trained model to predict some values on the balanced dataset and it resulted in the following confusion matrix:

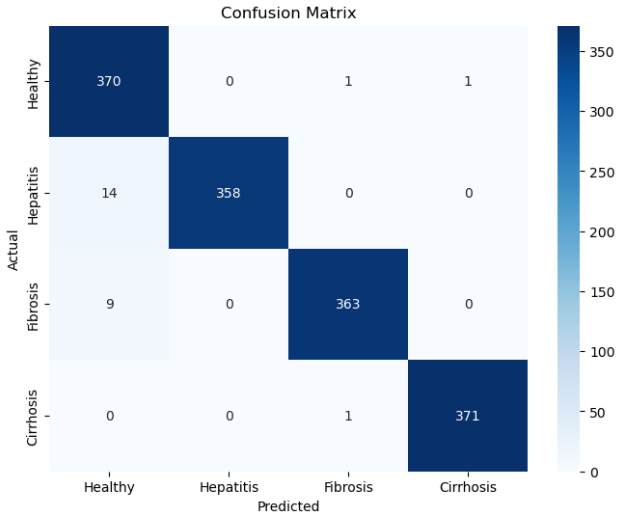

Fig.7 Model's Confusion Matrix

The confusion matrix shows that model's performance is extraordinary, as it achieved a **Percision score of 99%, Recall score of 99%**, and **F1-score of 99%** as well.

The final point to be discussed in this section is the SHAP values, as we had 12 features in the dataset, it was a very important insight to know which feature affected the final prediction the most and which feature's presence didn't immensely impact the final decision.

The following figure illustrates the SHAP values for each feature:

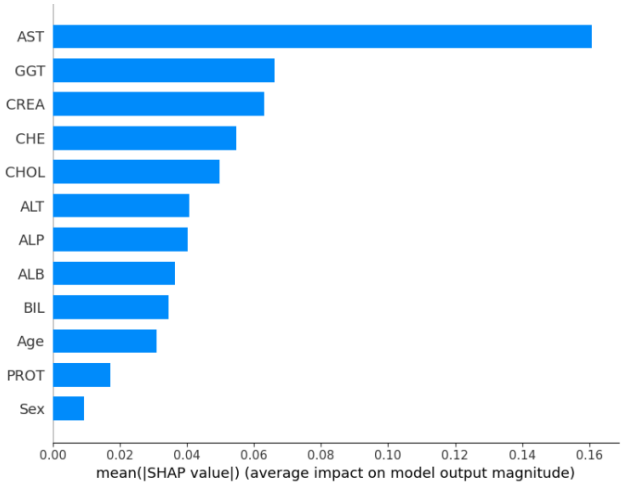

Fig.8 SHAP Values

VI. CONCLUSION

This study demonstrates the effectiveness of machine learning in early detection and classification of hepatitis C related liver diseases, especially cirrhosis and fibrosis by evaluating different algorithms, prioritization methods, prediction model a robust with outstanding performance specifications The development was enabled. The absence of early symptoms in chronic hepatitis C underscores the critical need for prompt diagnosis and early intervention. Our proposed solution, a predictive model based on machine learning algorithms, addresses the challenges associated with asymptomatic disease progression Using model training and fine-grained validation using the UCI Machine Learning Repository dataset of patient populations and laboratories.

The results reveal the successful use of SMOTE for balancing the dataset and the successful use of Optuna for hyperparameter tuning, which participate in the development of models with 97% validation accuracy and 95% test accuracy.

Relevant works in the liver disease prediction area exploration, ensures the Significance and potential impact of machine learning applications in medical diagnostics. Comparative studies with other models showcased the superiority of our approach further attests to the effectiveness of our proposed methodology.

By analyzing the SHAP values, the understanding which was got provided predictions of how every factor affected the model and helped greatly to understand the most importantly affecting factors depending on who was in which category. This can be used by the researchers and the clinicians in the future on how to better make their prescriptions.

The machine learning model designed in this research has great potential to enhance early detection and classification of the specific hepatitis C-related liver diseases with peculiar performance considerations and insights from the feature importance analysis allow saying that it is a valuable tool to improve medical research and contribute to improving public health interventions. As machine learning evolves, utilization of such models in clinical practice may truly be transformative to liver diseases' management by fostering care innovations implemented proactively for everybody.

VII. REFERENCES

[1] Kashif, A. A., Bakhtawar, B., Akhtar, A., Akhtar, S., Aziz, N., & Javeid, M. S. (2021). Treatment Response Prediction in Hepatitis C Patients using Machine Learning Techniques. International Journal of Technology, Innovation and Management (IJTIM), 1(2), 79–89. https://doi.org/10.54489/ijtim.v1i2.24

[2] S. Hashem et al., "Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 15, no. 3, pp. 861-868, 1 May-June 2018, doi: 10.1109/TCBB.2017.2690848.

[3] Konerman MA, Beste LA, Van T, Liu B, Zhang X, et al. (2019) Machine learning models to predict disease progression among veterans with hepatitis C virus. PLOS ONE 14(1): e0208141. https://doi.org/10.1371/journal.pone.0208141

[4] K. Ahammed, M. S. Satu, M. I. Khan and M. Whaiduzzaman, "Predicting Infectious State of Hepatitis C Virus Affected Patient's Applying Machine Learning Methods," 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 2020, pp. 1371-1374, doi: 10.1109/TENSYMP50017.2020.9230464.

[5] A. M. Hauri, G. L. Armstrong, and Y. J. F. Hutin, â€œThe global burden of disease attributable to contaminated injections given in health care settings,â€ Int. J. STD AIDS, vol. 15, no. 1, pp. 7â€“16, 2004, doi: 10.1258/095646204322637182.

[6] G. Hoffmann, A. Bietenbeck, R. Lichtinghagen, and F. Klawonn, â€œUsing machine learning techniques to generate laboratory diagnostic pathwaysâ€”a case study,â€ J. Lab. Precis. Med., vol. 3, pp. 58â€“58, 2018, doi: 10.21037/jlpm.2018.06.01.

[7] World Health Organization, “Hepatitis C [Hepatitis C]

[8] Stoean, R., Stoean, C., Lupsor, M., Stefanescu, H., & Badea, R. (2011). Evolutionary-driven support vector machines for determining the degree of liver fibrosis in chronic hepatitis C. Artificial intelligence in medicine, 51(1), 53-65.

[9] Mohamed, A. A., Elbedewy, T. A., El-Serafy, M., El-Toukhy, N., Ahmed, W., & El Din, Z. A. (2015). Hepatitis C virus: A global view. World journal of hepatology, 7(26), 2676.

[10] Liu, X., Xiao, Z., Song, Y., Zhang, R., Li, X., & Du, Z. (2021). A Machine Learning-Aided Framework to Predict Outcomes of Anti-PD-1 Therapy for Patients With Gynecological Cancer on Incomplete Post-Marketing Surveillance Dataset. IEEE Access, 9, 120464-120480.

[11] Saad, Y., Awad, A., Alakel, W., Doss, W., Awad, T., & Mabrouk, M. (2018). Data mining of routine laboratory tests can predict liver disease progression in Egyptian diabetic patients with hepatitis C virus (G4) infection: a cohort study of 71 806 patients. European journal of gastroenterology & hepatology, 30(2), 201-206.

[12] Baptista, D., Ferreira, P. G., & Rocha, M. (2021). Deep learning for drug response prediction in cancer. Briefings in bioinformatics, 22(1), 360-379.

[13] Ahmad, M., Aftab, S., Ali, I., & Hameed, N. (2017). Hybrid tools and techniques for sentiment analysis: a review. Int. J. Multidiscip. Sci. Eng, 8(3), 29-33.

[14] Chieh-Chen Wu a e, Wen-Chun Yeh b, Wen-Ding Hsu c, Md. Mohaimenul Islam a e, Phung Anh (Alex) Nguyen e, Tahmina Nasrin Poly a e,

Yao-Chin Wang a e d, Hsuan-Chia Yang e, Yu-Chuan (Jack) Li

[15] (S. Katiyar, “Predictive analysis on diabetes, liver and kidney diseases using machine learning,” International Journal for Research in Applied Science and Engineering Technology, vol. 8, no. 5, pp. 2285–2292, 2020.)