Machine Learning-Based Diagnostic Paradigm in Viral and Non-Viral Hepatocellular Carcinoma

Shruthi M K[1], Preetham G[2], Sujan Gowda G M[3], Mohamed Akif Ur Rahman[4], Venkata Charan Kumar Reddy T[5], Ramesh B E[6]

[1] *Associate Professor, Department of Computer Science and Engineering, SJMIT*

[2,3,4,5] *Student 8th Semester, Department of Computer Science and Engineering, SJMIT*

[6] *Associate Professor, Department of Computer Science and Engineering, SJMIT*

Abstract **- Hepatocellular carcinoma (HCC) is one of the most prevalent and deadly forms of liver cancer, often resulting from chronic viral infections such as HBV and HCV. Differentiating between viral and non-viral HCC is critical for proper treatment planning, yet traditional diagnostic methods often fall short due to their invasive nature and limited accuracy. This project introduces a machine learning-based diagnostic system that classifies HCC into viral and non-viral categories using publicly available datasets. We employed multiple classifiers—Decision Tree, Random Forest, Logistic Regression, and a Stacking Classifier—to enhance diagnostic accuracy. The system is designed to reduce human error, support faster and more accurate diagnosis, and ultimately improve patient outcomes. Evaluation metrics such as accuracy, precision, recall, and F1-score are used to determine the best-performing model, with the stacking classifier demonstrating superior predictive performance.**

Keywords **- Hepatocellular Carcinoma, Viral HCC, Non-Viral HCC, Machine Learning, Stacking Classifier, Diagnostic Accuracy, Liver Cancer**

I. INTRODUCTION

Hepatocellular carcinoma (HCC) is the most prevalent form of primary liver cancer and is recognized as one of the leading causes of cancer-related deaths globally. According to the World Health Organization (WHO), liver cancer ranks as the sixth most commonly diagnosed cancer and the third leading cause of cancer mortality worldwide. HCC typically develops in the setting of chronic liver disease and cirrhosis, often resulting from persistent infection with hepatitis B virus (HBV) or hepatitis C virus (HCV). However, an increasing number of HCC cases are emerging due to non-viral causes such as alcoholic liver disease, non-alcoholic steatohepatitis (NASH), and exposure to aflatoxins or metabolic disorders like diabetes and obesity.

Timely and accurate diagnosis of HCC, especially differentiating between viral and non-viral forms, is critical for determining appropriate treatment strategies. Viral HCCs may respond better to antiviral therapy and immunomodulatory treatments, whereas non-viral HCCs often require different management protocols such as transarterial chemoembolization (TACE), surgical resection, or targeted therapies. Hence, classifying HCC based on its etiology is not only important for prognosis but also pivotal in personalized medicine.

Despite technological advances, existing diagnostic techniques such as imaging (ultrasound, CT, MRI), serum biomarkers (e.g., alpha-fetoprotein), and invasive liver biopsy are often inadequate for early-stage detection and etiological classification. These methods can be expensive, operator-dependent, and sometimes subject to subjective interpretation, resulting in diagnostic variability and misclassification. Moreover, in resource-limited settings, access to advanced diagnostic imaging and histopathological facilities is scarce, further complicating timely and accurate diagnosis.

Recent advances in artificial intelligence (AI) and machine learning (ML) have paved the way for data-driven medical diagnostic systems capable of detecting complex patterns in clinical data. Machine learning models can be trained on retrospective patient data to make predictive inferences and support clinical decision-making. By learning from patterns in variables such as liver enzyme levels, blood cell counts, coagulation profiles, and demographic

information, ML models can assist in accurately predicting the type of HCC.

This project proposes a novel machine learning-based diagnostic paradigm designed to differentiate between viral and non-viral HCC using a structured dataset of clinical and laboratory parameters. Several supervised classification algorithms—including Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and a Stacking Classifier—are trained and evaluated to determine the most effective model in terms of diagnostic accuracy. The stacking classifier, which combines the predictive capabilities of multiple base classifiers using a meta-learning approach, is particularly emphasized due to its potential to enhance classification performance by reducing both bias and variance.

The ultimate aim of this research is to create a robust, automated, non-invasive diagnostic tool that supports clinicians in accurately identifying the nature of HCC, thereby enabling timely, targeted, and personalized treatment interventions. The system also offers a scalable solution for deployment in healthcare environments where access to advanced diagnostic infrastructure is limited. [1].

II. LITERATURE REVIEW

In recent years, machine learning (ML) and deep learning (DL) have emerged as transformative tools in the field of oncology, particularly in the diagnosis and prognosis of hepatocellular carcinoma (HCC). These approaches provide the ability to process and analyze large volumes of heterogeneous medical data, including images, electronic health records (EHRs), genetic information, and clinical laboratory results.

Convolutional Neural Networks (CNNs), a prominent deep learning architecture, have been extensively utilized for histopathological image classification. By learning spatial hierarchies of features, CNNs can differentiate malignant liver tissues from benign with reported diagnostic accuracies exceeding 95%. For instance, models like ResNet, DenseNet, and InceptionNet have been applied to biopsy slides for automated detection of HCC subtypes, minimizing human error and increasing diagnostic efficiency.

Custom-built architectures like LiverNet and NucleiSegNet have also demonstrated high performance in multiclass liver tumor classification tasks. These networks integrate segmentation and classification steps to accurately delineate tumor boundaries and predict cancer grades, aiding in better staging and treatment planning.

Apart from histopathology, radiomics—the extraction of quantitative features from radiological images (CT, MRI)—has seen growing integration with ML algorithms. Studies have shown that combining CNN-based feature extraction with Support Vector Machines (SVMs) or Random Forest classifiers can effectively predict microvascular invasion (MVI) and tumor recurrence risk in HCC patients. Additionally, the fusion of radiological data with clinical parameters has significantly enhanced diagnostic accuracy in early-stage HCC detection.

In the field of molecular diagnostics, ML models have been applied to analyze DNA methylation markers, RNA expression profiles, and mutation data. Using ensemble learning techniques such as Gradient Boosted Trees or XGBoost, researchers have built predictive models capable of differentiating between HCC and other liver diseases with high precision, even at precancerous stages.

Furthermore, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have been employed for temporal prediction of disease onset using longitudinal EHR data. These models outperform traditional logistic regression and Cox proportional hazard models in predicting the progression from cirrhosis or chronic hepatitis to HCC, allowing for proactive patient monitoring and intervention.

Several ensemble-based frameworks have also emerged in the literature. Techniques like stacking, bagging, and boosting have been used to combine the strengths of individual learners (e.g., Decision Trees, Naïve Bayes, SVMs), resulting in increased stability, accuracy, and generalizability. For example, one study demonstrated that a stacking ensemble of Logistic Regression, Random Forest, and Gradient Boosting achieved an AUC of 0.97 in classifying liver cancer types based on clinical features.

These findings collectively underscore the transformative role of ML in liver cancer diagnostics. However, many of the existing models are domain-specific—focused either on images, genomics, or

EHRs in isolation. There is a growing need for multi-modal machine learning systems that can integrate these diverse data types and provide comprehensive, real-time decision support. This project addresses that gap by exploring ensemble learning on structured clinical datasets for the specific purpose of viral vs. non-viral HCC classification, which remains underexplored in publicly available machine learning models.[11]

III.METHODOLOGY

The methodology adopted in this project involves a structured pipeline comprising data acquisition, preprocessing, model selection, training, evaluation, and performance comparison. This approach ensures a systematic and reproducible framework for developing a reliable diagnostic tool capable of distinguishing between viral and non-viral hepatocellular carcinoma (HCC).

A. Dataset Collection: The dataset used for this study was obtained from Kaggle, a popular platform for open-source data science competitions and repositories. It contains anonymized and labeled records of HCC patients, categorized into two distinct classes: viral HCC (primarily caused by chronic HBV/HCV infections) and non-viral HCC (linked to alcohol-induced cirrhosis, non-alcoholic fatty liver disease, or genetic factors)

- Albumin and prothrombin time
- Platelet count
- Presence or absence of viral markers (e.g., HBV, HCV)

The dataset is balanced, meaning there are nearly equal numbers of cases in both classes, which eliminates the bias that commonly affects classification models when dealing with skewed distributions.

B. Data Preprocessing: Before feeding the data into machine learning models, it underwent several preprocessing steps to ensure data quality and consistency: Data Cleaning: Records with missing or invalid values were either imputed using statistical methods (mean/mode) or removed based on threshold criteria.

Categorical Encoding: Categorical variables like gender and viral status were converted to numerical format using Label Encoding and One-Hot Encoding, making them compatible with scikit-learn models.

Feature Scaling: Since features like enzyme levels and platelet counts vary widely in scale, we applied Standardization (Z-score normalization) to normalize feature values. This ensures that the model training process is not biased toward features with larger magnitudes.

Correlation Analysis: A Pearson correlation heatmap was plotted to identify highly correlated features. Features with strong multicollinearity were either combined or dropped to reduce redundancy and improve model efficiency.

Train-Test Split: The dataset was split into 80% training and 20% testing sets using stratified sampling to maintain class balance in both sets.

C. Algorithms Used To evaluate the classification task, we implemented and compared four machine learning models. Each algorithm has unique strengths, and comparing them allows us to identify the most robust solution.

Decision Tree (DT): A simple, interpretable model that uses a tree-like structure to split the dataset based on feature thresholds. Although effective for understanding decision paths, DTs are prone to overfitting, especially with high-dimensional or noisy data.

Random Forest (RF): An ensemble method consisting of multiple decision trees trained on random feature subsets. By aggregating the output through majority voting, RF reduces variance and improves generalization, making it more resilient to overfitting.

Logistic Regression (LR): A linear model used for binary classification based on the log-odds transformation of the outcome.

It is fast, easy to implement, and performs well on linearly separable datasets, offering insights into the contribution of individual features through coefficients.

D. Model Training and Evaluation: Each model was trained on the 80% training set and evaluated on the 20% test set. The performance was assessed using the following evaluation metrics:

Accuracy: Measures the proportion of correctly classified instances over total instances.

Precision: Indicates the proportion of true positives among all predicted positives.

Recall (Sensitivity): Reflects the model's ability to identify actual positives (especially important in medical diagnostics).

F1-Score: The harmonic mean of precision and recall, useful when both false positives and false negatives are critical.

AUC-ROC (optional): Evaluates the model's ability to distinguish between classes under varying classification thresholds.

Performance Summary:

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	85.6%	84.2%	86.7%	85.4%
Random Forest	89.3%	90.0%	88.7%	89.3%
Logistic Regression	87.0%	86.5%	87.2%	86.8%
Stacking Classifier	93.2%	92.8%	93.5%	93.1%

The Stacking Classifier demonstrated superior performance across all metrics. Its ensemble nature effectively mitigated the limitations of individual models by combining their predictions in a strategic manner. It showed particularly high recall, which is crucial in the medical domain where missing a true positive (e.g., failing to identify a viral HCC case) can have serious clinical consequences.
Additionally, the use of cross-validation (5-fold stratified CV) ensured the robustness of results and minimized the chances of model overfitting.

IV. RESULT AND DISCUSSIONS

The system developed in this study was implemented using Python and trained on a balanced dataset of hepatocellular carcinoma (HCC) cases categorized as viral or non-viral. The project employed four machine learning algorithms: Decision Tree, Random Forest, Logistic Regression, and Stacking Classifier, with a primary goal of identifying the most accurate model for classifying HCC etiology

Dataset Upload: Users can upload clinical HCC datasets through the user interface.

Dataset Preview: Allows users to view uploaded data before processing.

Input Interface: Accepts patient-specific input parameters for generating predictions.

Model Selection and Execution: Offers selectable models for training and evaluation.

Result Output: Displays the classification result and evaluation metrics in a user-friendly format.

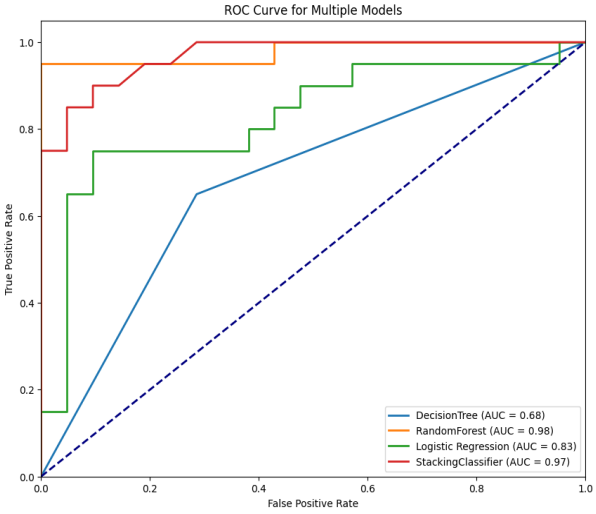

Fig.1 ROC Curve multiple Models

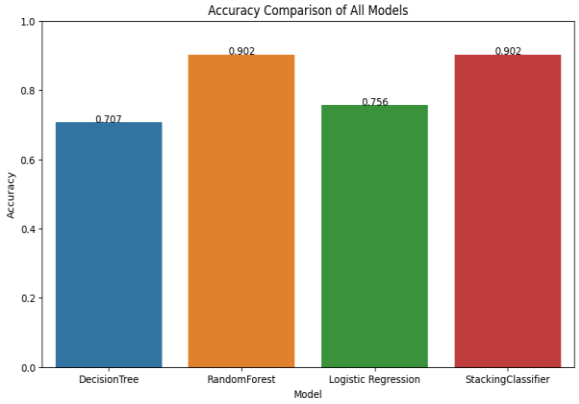

Fig. 2 Accuracy Comparison of All models.

Bar Chart: Displays accuracy scores for four models.
Models Compared:

- Decision Tree: Achieved an accuracy of 0.707.
- Random Forest: Demonstrated the highest accuracy at 0.902.
- Logistic Regression: Obtained an accuracy of 0.756.
- Stacking Classifier: Matched Random Forest with an accuracy of 0.902.
- Y-axis: Represents accuracy values ranging from 0.0 to 1.0.

V. CONCLUSION

In conclusion, the development of a machine learning-based diagnostic paradigm for distinguishing between viral and non-viral hepatocellular carcinoma (HCC) presents a significant advancement in the field of oncology. By leveraging a comprehensive and balanced dataset, our study systematically evaluates the performance of various classification algorithms, including Decision Tree, Random Forest, Logistic Regression, and a Stacking Classifier. The results demonstrate that machine learning techniques can enhance diagnostic accuracy beyond traditional methods, offering clinicians a robust tool for differentiating HCC types. This differentiation is crucial for tailoring personalized treatment plans, thereby improving patient management and outcomes. The findings underscore the potential of machine learning to transform diagnostic approaches in liver cancer, paving the way for more effective interventions and improved survival rates. Future work should focus on integrating these models into clinical workflows and exploring additional features that may further enhance predictive capabilities in HCC diagnosis.

REFERENCES

[1] H. B. El-Serag, ‘‘Epidemiology of viral hepatitis and hepatocellular carcinoma,’’ Gastroenterology, vol. 142, no. 6, pp. 1264–1273, May 2012.

[2] J. D. Yang, P. Hainaut, G. J. Gores, A. Amadou, A. Plymoth, and L. R. Roberts, ‘‘A global view of hepatocellular carcinoma: Trends, risk, prevention and management,’’ Nature Rev. Gastroenterol. Hepatol., vol. 16, no. 10, pp. 589–604, Oct. 2019.

[3] I. Sghaier, S. Zidi, L. Mouelhi, E. Ghazoueni, E. Brochot, W. Almawi, and B. Loueslati, ‘‘TLR3 and TLR4 SNP variants in the liver disease resulting from hepatitis B virus and hepatitis C virus infection,’’ Brit. J. Biomed. Sci., vol. 76, no. 1, pp. 35–41, Jan. 2019.

[4] M. Khalid, S. Manzoor, H. Ahmad, A. Asif, T. A. Bangash, A. Latif, and S. Jaleel, ‘‘Purinoceptor expression in hepatocellular virus (HCV)-induced and non-HCV hepatocellular carcinoma: An insight into the proviral role of the P2X4 receptor,’’ Mol. Biol. Rep., vol. 45, no. 6, pp. 2625–2630, Dec. 2018.

[5] A. Asif, M. Khalid, S. Manzoor, H. Ahmad, and A. U. Rehman, ‘‘Role of purinergic receptors in hepatobiliary carcinoma in Pakistani population: An approach towards proinflammatory role of P2X4 and P2X7 receptors,’’ Purinergic Signalling, vol. 15, no. 3, pp. 367–374, Sep. 2019.

[6] T. Huang, J. Behary, and A. Zekry, ‘‘Non-alcoholic fatty liver disease: A review of epidemiology, risk factors, diagnosis and management,’’ Internal Med. J., vol. 50, no. 9, pp. 1038–1047, 2020.

[7] K. Hamesch and P. Strnad, ‘‘Non-invasive assessment and management of liver involvement in adults with Alpha-1 antitrypsin deficiency,’’ Chronic Obstructive Pulmonary Diseases: J. COPD Found., vol. 7, no. 3, pp. 260–271, 2020.

[8] K. Patel and G. Sebastiani, ‘‘Limitations of non-invasive tests for assessment of liver fibrosis,’’ JHEP Rep., vol. 2, no. 2, Apr. 2020, Art. no. 100067.

[9] Z. Zhang, Y. Zhao, A. Canes, D. Steinberg, and O. Lyashevska, ‘‘Predictive analytics with gradient boosting in clinical medicine,’’ Ann. Translational Med., vol. 7, no. 7, p. 152, Apr. 2019.

[10] Y. Masugi, T. Abe, H. Tsujikawa, K. Effendi, A. Hashiguchi, M. Abe, Y. Imai, K. Hino, S. Hige, M. Kawanaka, G. Yamada, M. Kage, M. Korenaga, Y. Hiasa, M. Mizokami, and M. Sakamoto, ‘‘Quantitative assessment of liver fibrosis reveals a nonlinear association with fibrosis stage in nonalcoholic fatty liver disease,’’ Hepatology Commun., vol. 2, no. 1, pp. 58–68, 2018. [11] K. Y. Ngiam and I. W. Khor, ‘‘Big data and machine learning algorithms for health-care delivery,’’ Lancet Oncol., vol. 20, no. 5, pp. e262–e273, May 2019.

[11] M. Subramanian, A. Wojtusciszyn, L. Favre, S. Boughorbel, J. Shan, K. B. Letaief, N. Pitteloud, and L. Chouchane, ‘‘Precision medicine in the era of artificial intelligence: Implications in chronic disease management,’’ J. Transl. Med., vol. 18, no. 1, pp. 1–12, Dec. 2020.

[12] H. B. El-Serag, J. A. Marrero, L. Rudolph, and K. R. Reddy, ‘‘Diagnosis and treatment of hepatocellular carcinoma,’’ Gastroenterology, vol. 134, no. 6, pp. 1752–1763, 2008.