IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 5, Issue 5, April 2025

Impact Factor: 7.67

Heptocellular Carcinoma (HCC) Detection using Machine Learning

B. Sri Pavani[1], SK. Sameena[1], K. Bhuvana Sai Kalyani[1], E. Bala Barath[1], P. Rajesh[2]

Students, Department of Computer Science and Engineering[1]

Associate Professor, Department of Computer Science and Engineering[2]

SRK Institute of Technology, NTR, Andhra Pradesh, India

Abstract: *Hepatocellular carcinoma (HCC) is a common and aggressive form of liver cancer, and accurate diagnosis is critical for effective treatment and patient management. This project focuses on developing a machine learning-based diagnostic paradigm to distinguish between viral and non-viral HCC cases using a comprehensive dataset from Kaggle. The dataset, which includes balanced cases of HCC, serves as the foundation for our analysis. We employ several classification algorithms to achieve this, including Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and a Stacking Classifier. Each algorithm is evaluated for its performance in accurately classifying the type of HCC, with the goal of identifying the most effective method for diagnosis. The Stacking Classifier, which combines multiple models to improve predictive accuracy, is of particular interest in this study. By comparing the results of these models, we aim to enhance diagnostic precision, support personalized treatment plans, and ultimately contribute to better patient outcomes. This project seeks to address the limitations of traditional diagnostic methods and provide a robust tool for clinicians to differentiate between viral and non-viral HCC effectively.*

Keywords*:* Hepatocellular carcinoma

I. INTRODUCTION

Hepatocellular carcinoma (HCC) is the most prevalent form of primary liver cancer and ranks among the leading causes of cancer-related deaths globally. It primarily arises in individuals with chronic liver diseases such as hepatitis B virus (HBV), hepatitis C virus (HCV), cirrhosis, and non-alcoholic fatty liver disease (NAFLD). The aggressive progression of HCC, combined with its asymptomatic nature in the early stages, often results in delayed diagnosis and limited treatment options. This delay is critical, as early detection plays a vital role in determining prognosis and therapeutic effectiveness.

Traditional diagnostic techniques for HCC, including imaging methods like ultrasound, computed tomography (CT), and magnetic resonance imaging (MRI), along with liver biopsies, have limitations in accurately identifying and classifying HCC.

These approaches can be invasive, expensive, and sometimes inconclusive, especially when attempting to distinguish HCC from other liver conditions. Moreover, misdiagnosis remains a persistent challenge, leading to inappropriate treatment strategies and poor patient outcomes. Given these constraints, enhancing diagnostic precision has become an area of urgent focus in liver cancer research.

One of the most critical challenges in HCC management is the need to differentiate between **viral and non-viral etiologies**. This classification is essential, as the underlying cause significantly influences treatment decisions and prognosis. Viral HCC, often linked to chronic HBV or HCV infections, differs in pathophysiology and response to therapy when compared to non-viral HCC, which may stem from metabolic disorders, alcohol abuse, or NAFLD. Misclassifying the disease can hinder personalized treatment planning, reducing the chances of successful outcomes. Therefore, a more evidence-based decisions.

In recent years, **machine learning (ML)** has emerged as a promising solution to the limitations of traditional diagnostic methods. ML models can analyze complex datasets and uncover subtle patterns and correlations that may not be

DOI: 10.48175/IJARSCT-25242

apparent to human observers. These capabilities are particularly valuable in medical diagnostics, where data-driven decision-making can significantly improve diagnostic accuracy. This project explores the application of multiple ML algorithms—**Decision Tree, Random Forest, Logistic Regression**, and a **Stacking Classifier**—to classify HCC cases into viral and non-viral categories using a balanced and comprehensive dataset.

The core objective of this research is to develop a reliable and accurate ML-based diagnostic tool that enhances the classification of HCC etiologies. By evaluating the performance of different algorithms and comparing their diagnostic capabilities, the study aims to identify the most effective method for distinguishing between viral and non-viral HCC. This, in turn, can support the development of personalized treatment strategies, reduce misclassification rates, and contribute to improved clinical outcomes.

Ultimately, this project highlights the growing potential of AI-driven solutions in modern healthcare. By addressing the limitations of conventional diagnostic approaches, it not only supports timely and precise identification of HCC types but also lays the foundation for more targeted and effective cancer care. The integration of ML in liver cancer diagnostics could transform how physicians manage the disease, ensuring better patient outcomes and more efficient use of healthcare resources.

II. LITERATURE REVIEW

In recent years, a considerable amount of research has been conducted on managing hepatocellular carcinoma (HCC) because it is challenging to manage the rising worldwide burden of HCC and its complex association with chronic liver diseases. Early defining contributions included a review by El-Serag (2012)[1], highlighting the mounting significance of chronic hepatitis B and C virus infections as potential risk factors for HCC.This study focused on the epidemiological aspects of viral hepatitis, describing the global.

In a complementary trajectory, Khalid et al. Here we report the study presenting molecular mechanisms that drive HCC, with attention to purinoceptor expression in hepatitis C virus (HCV) positive and negative HCCs (2018)[2]. They found the P2X4 receptor was upregulated in carcinomas induced by HCV, while P2X7 levels remained stable, hinting at a provirial role of P2X4 in hepatic tumorigenesis. This molecular insight additionally provides crucial layer of our comprehension of HCC pathogenesis at the receptor level.

Drawing on the broader global landscape of biliary tract cancer, Yang et al. In 2019[3], there emerged a broad look at big-picture trends, risk factors that get people into trouble, prevention strategies folks have used and management methods that work. Their study illuminated HCC as one of the leading causes of cancer-related mortality worldwide, advocating for early detection protocols and improved therapeutic frameworks to alleviate disease burden. These results really point to the absolute need for a strategic response to HCC that's global in scope.

Sghaier and colleagues dug deeper into how genetics and immune stuff play into HCC too. In 2019[4], someone looked at how certain genetic variations, called Single Nucleotide Polymorphisms or SNPs, affected how well the liver progressed when not healthy. These variations relate specifically to something called Toll Like Receptors or TLRs specifically TLR 3 and TLR 4. Studying these receptors can give us important clues about how the sensitive organ cares for itself and thrives or when it gets sick and starts to deteriorate. Their research identified certain markers that are linked to higher risk for cirrhosis and liver cancer in people who have chronic infection from Hepatitis B and HCV. These markers also have potential to be useful for predicting risk and providing early detection.

In a more technologically forward-looking perspective, Subramanian et al. (2020) [5] examined the role of artificial intelligence and high-performance computing in transforming chronic disease management. Their research promoted precision medicine for diagnosing and treating liver inflammation because they emphasize customizing interventions that really address individual patients differently. It’s like solving a medical case based on who the patient is and what makes that person's liver inflamed. It's really focused and very tailored to the specific individual and personal health situation. This work captures really nicely the recent very cool blend of computing improvements and practice in HCC medicine.

As a whole, these studies provide a comprehensive understanding of hepatocellular carcinoma from an epidemiological, molecular, genetic, and technological viewpoint. This literature survey compiles important findings gained from recent studies to serve as a sound basis for further research aimed at the diagnosis, treatment, and prevention of HCC.

III. EXISTING SYSTEM

Currently,relevant imaging modalities such as ultrasound, CT and MRI, alongside biopsy histopathology and clinical examination, make up the composite diagnostic criteria for hepatocellular carcinoma (HCC). These traditional methods are essential to the diagnosis of HCC, but have relative shortcomings, particularly in differentiating viral (HBV and HCV)-versus non-viral factors and causes.

- **Subjective Interpretation**: Assessing the accuracy of a diagnosis highly relies on the experience and judgment of radiologists and pathologists, creating variation and possible bias in the outcome.
- **Inter-Method and Inter-Clinician Variability**: The different types of imaging techniques or diagnostic methods used, together with variability between clinicians, may result in divergent and contradictory findings.
- **Risk of Misclassification:** The overlapping clinical features associated with both non-viral and viral HCC could result in misdiagnosis and deviation in the optimal or correct treatment plan.
- **Limited Diagnostic Precision:** More traditional approaches are oftentimes unable posess the level of detail and sensitivity needed to clearly identify the cause of HCC, particularly in its earlier stages or whenever the situation is ambiguous.
- **Overlapping Clinical Features:** Viral and non-viral HCC can present with similar imaging and clinical characteristics, makingit difficult to achieve a cleardifferential diagnosis using standard procedures.

IV. PROPOSED SYSTEM

Proposed system strongly aims to improve HCC diagnosis by using machine learning algorithms for more accurate classification of hepatocellular carcinoma, either being viral or non-viral. The model utilizes a balanced dataset from Kaggle and uses Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and a Stacking Classifier for evaluation and comparison purposes. This process includes utilizing data preprocessing techniques, forcing the models, and then evaluating them through metrics including accuracy, precision, recall, and F1-score.The final output will be a robust classification tool that integrates the best-performing model, providing clinicians with a reliable and precise method for differentiating between viral and non-viral HCC cases. By automating the diagnostic process, the system aims to reduce human error, improve diagnostic accuracy, and facilitate personalized treatment strategies for patients.

- **Improved Accuracy**: The machine learning algorithms classify HCC types more accurately than conventional methods.
- **Minimized Human Error:** Automated diagnosis eliminates the risk of subjective interpretation errors.
- **Streamlined Workflow**: Improved efficiency in processing and classification of HCC cases.
- Last but not the least, machine learning models provide consistent results over various cases and datasets.
- **Subtype-Specific Therapy**: Enhanced diagnostic precision aids in the definition of subtype-targeted therapeutics by differentiating between varied HCC subtypes.

V. SYSTEM ARCHITECUTURE AND METHODOLOGY

In this section, a comprehensive description on the system architecture and methodological workflow constructed for classifying the data as viral or non-viral is delineated. The proposed work consists of a front end for user interaction and a back end which implements machine learning, so as to allow for efficient prediction on how well the user-poem pairs will do.

1. User Interface and Access Control

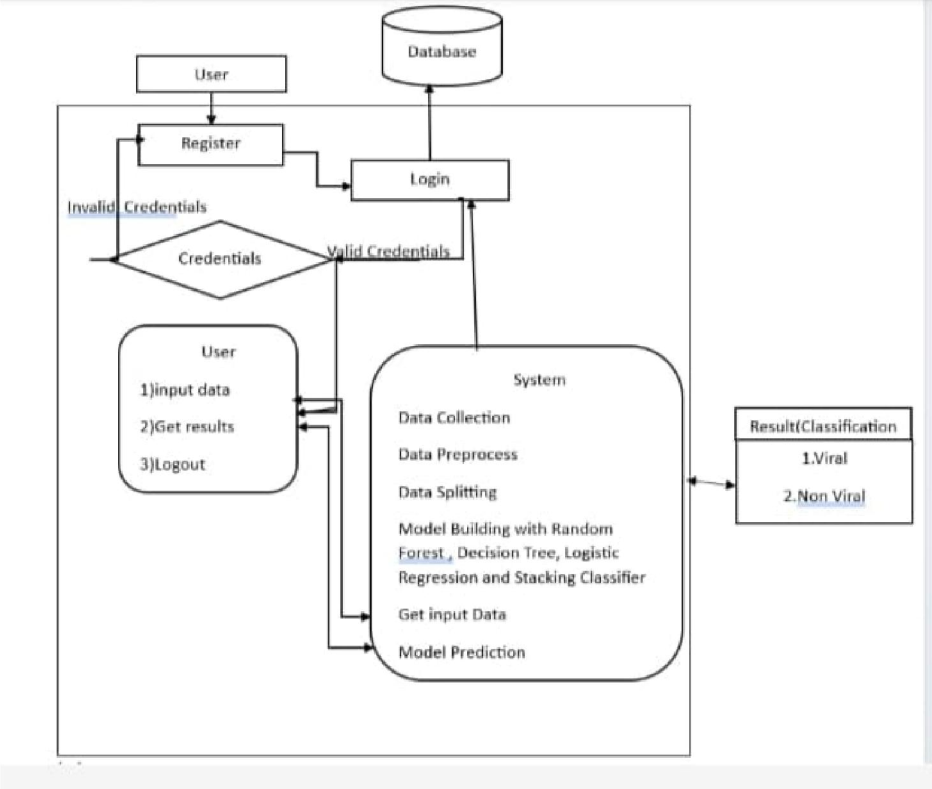
The system kicks off with a user-friendly interface that offers options for user registration and login. New users need to register, and their credentials are safely stored in a centralized database For those returning, logging in is as simple as entering their registered details.

When users submit their credentials:

- If the credentials are valid, they gain access to the main dashboard of the system.
- If the credentials are invalid, access is denied, and the user is prompted to enter the correct information again.

DOI: 10.48175/IJARSCT-25242

This setup ensures secure, role-based access and keeps unauthorized users at bay.



2. Functional Workflow

After logging in, users can access three key functionalities:

1. Input Data
2. Obtain Results
3. Logout

The backend of the system takes care of all data processing and model inference tasks once the data is submitted.

3. Machine Learning Pipeline

At the heart of the system lies a powerful machine learning pipeline, which consists of several important modules:

3.1 Data Collection

This module is in charge of gathering raw datasets from both external and internal sources. This data forms the backbone for training and evaluating the model.

3.2 Data Preprocessing

The raw data undergoes cleaning and transformation to ensure it meets quality and consistency standards. Typical preprocessing steps include:

- Removing missing or irrelevant values
- Normalizing and scaling the data
- Encoding categorical features
- Selecting or extracting features

DOI: 10.48175/IJARSCT-25242

3.3 Data Splitting:

The preprocessed dataset is divided into training and testing subsets. This division allows for unbiased evaluation of the model and ensures it can generalize well to new, unseen data.

3.4 Model Building:

The system uses an ensemble learning approach to boost its predictive performance. It trains several machine learning models, including:

- **Random Forest**: This is a collection of decision trees that utilizes bagging to enhance accuracy and minimize overfitting.
- **Decision Tree**: A model that organizes data by making splits based on information gain.
- **Logistic Regression**: A statistical model that works well for tasks involving binary classification.
- **Stacking Classifier**: This is a meta-model that merges predictions from various base classifiers to improve both accuracy and robustness.

These models are trained on the training dataset and then evaluated on the test set, focusing on performance metrics like accuracy, precision, recall, and F1-score.

3.5 Input Data and Prediction:

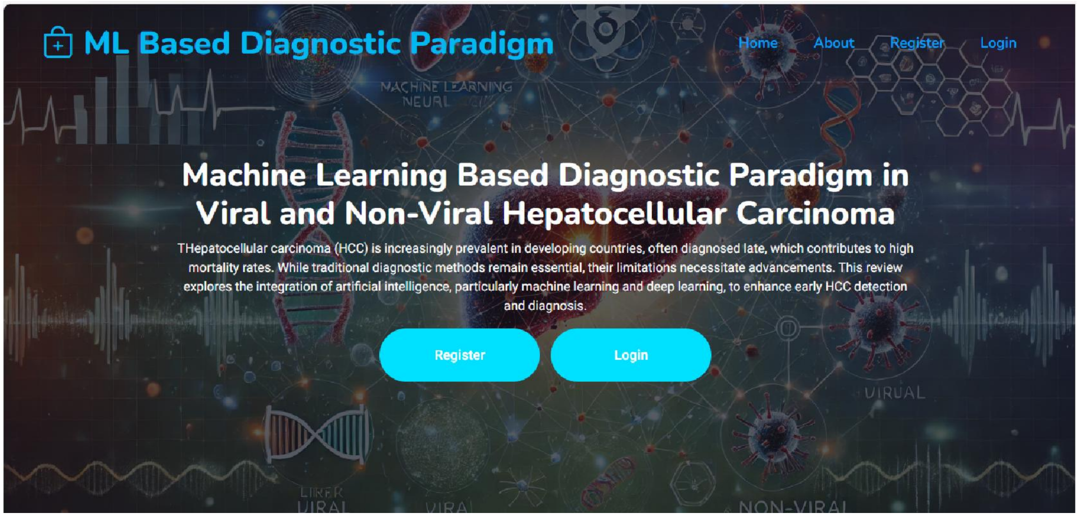
After the models are trained, users can enter new data through the interface. The system takes this input, processes it, and sends it to the trained model for prediction.

IV. RESULT GENERATION

Using the processed input and the trained model, the system generates a classification result, which is then shown to the user. The prediction can fall into one of these categories:

- Viral
- Non-Viral

V. RESULTS



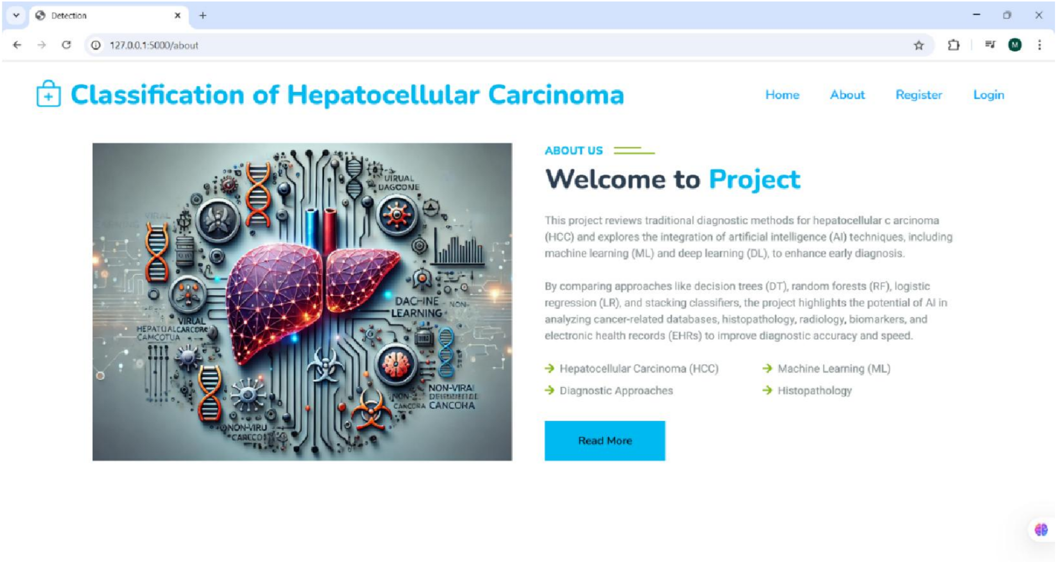
Machine Learning Based Diagnostic Paradigm in Viral and Non-Viral Hepatocellular Carcinoma

THepatocellular carcinoma (HCC) is increasingly prevalent in developing countries, often diagnosed late, which contributes to high mortality rates. While traditional diagnostic methods remain essential, their limitations necessitate advancements. This review explores the integration of artificial intelligence, particularly machine learning and deep learning, to enhance early HCC detection and diagnosis.

This interface enables users to, facilitating Register and login to model home page and prediction and training page

DOI: 10.48175/IJARSCT-25242

ABOUT PAGE:



Classification of Hepatocellular Carcinoma

ABOUT US

Welcome to Project

This project reviews traditional diagnostic methods for hepatocellular c arcinoma (HCC) and explores the integration of artificial intelligence (AI) techniques, including machine learning (ML) and deep learning (DL), to enhance early diagnosis.

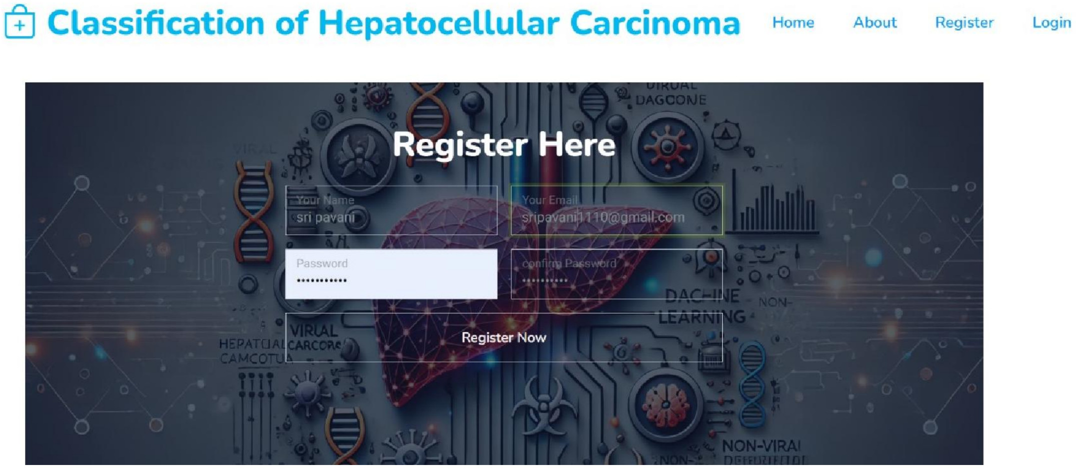
By comparing approaches like decision trees (DT), random forests (RF), logistic regression (LR), and stacking classifiers, the project highlights the potential of AI in analyzing cancer-related databases, histopathology, radiology, biomarkers, and electronic health records (EHRs) to improve diagnostic accuracy and speed.

→ Hepatocellular Carcinoma (HCC) → Machine Learning (ML)
→ Diagnostic Approaches → Histopathology

Fig:1About Page

The project predicts using machine learning models, emphasizing interpretability and superior accuracy with ensemble methods.

REGISTRATION:



Classification of Hepatocellular Carcinoma

Register Here

Your Name: sri pavani
Your Email: sripavani1110@gmail.com
Password: •••••••••••
confirm Password: ••••••••••

Fig:2 Registration Page

This page allows users to register for services, ensuring secure access by requiring personal details and password confirmation. It provides a user-friendly interface for creating a secure account.

DOI: 10.48175/IJARSCT-25242

LOGIN PAGE:

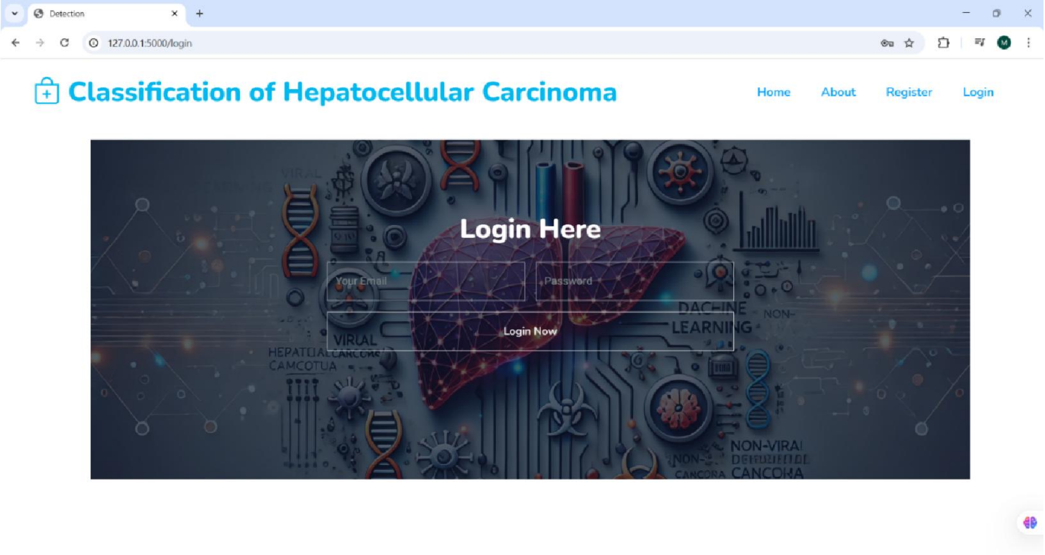
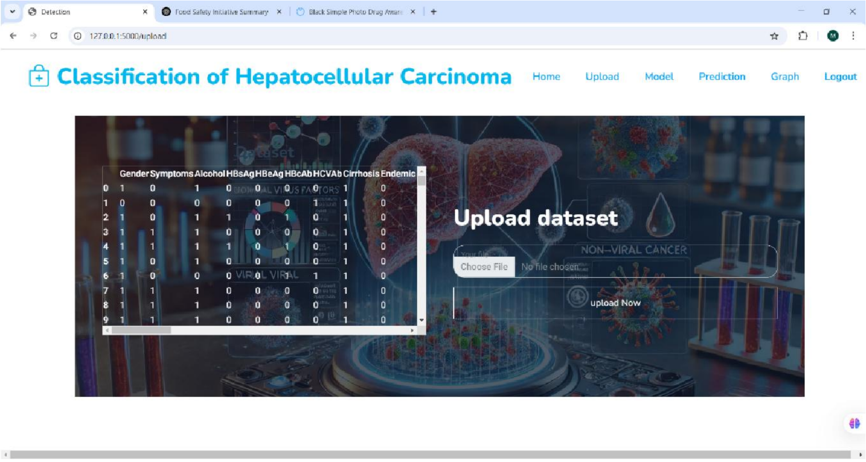


Fig:3 Login Page

This page provides a secure login interface for users to access the prediction account using their email and password.

UPLOAD PAGE



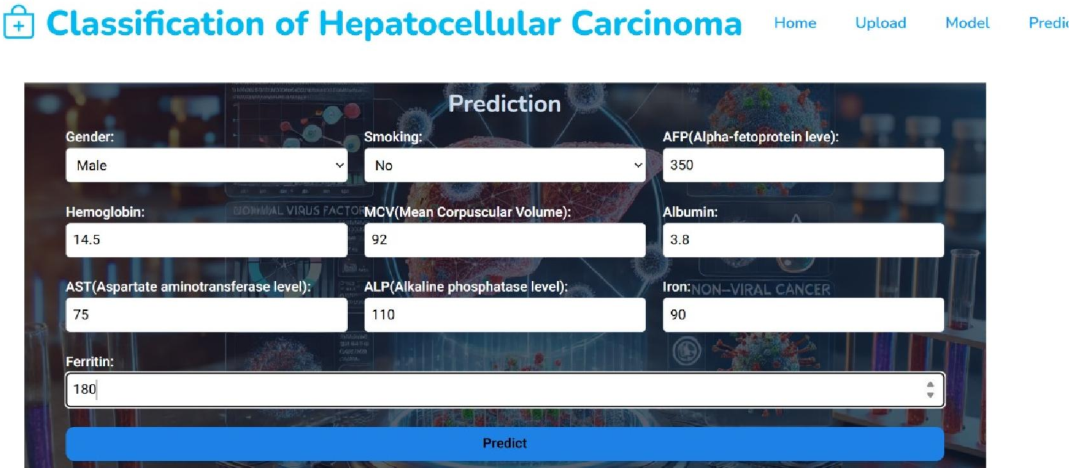
	Gender	Symptoms	Alcohol	HBsAg	HBeAg	HBcAb	HCVAb	Cirrhosis	Endemic
0	1	0	1	0	0	0	0	1	0
1	0	0	0	0	0	0	1	1	0
2	1	0	1	1	0	1	0	1	0
3	1	1	1	0	0	0	0	1	0
4	1	1	1	1	0	1	0	1	0
5	1	0	1	0	0	0	0	1	0
6	1	0	0	0	0	1	1	1	0
7	1	1	1	0	0	0	0	1	0
8	1	1	1	0	0	0	0	1	0
9	1	1	1	0	0	0	0	1	0

Fig:4 Upload Page

This page allows users to upload datasets for prediction, enabling model training and evaluation for accurate results.

DOI: 10.48175/IJARSCT-25242

PREDICTION PAGE:



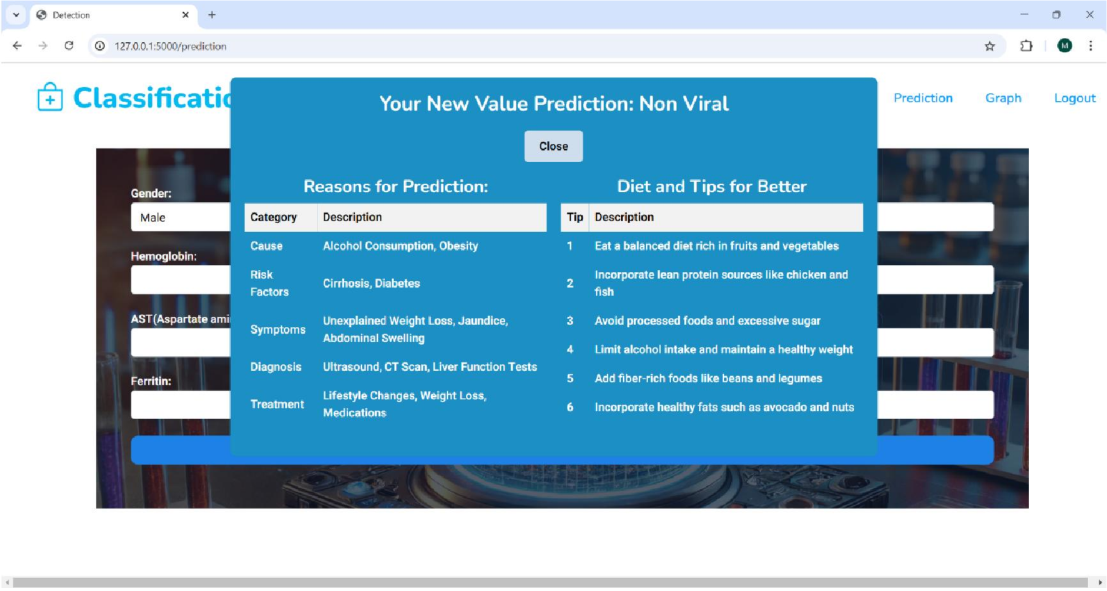
Prediction

Gender: Male	Smoking: No	AFP(Alpha-fetoprotein leve): 350
Hemoglobin: 14.5	MCV(Mean Corpuscular Volume): 92	Albumin: 3.8
AST(Aspartate aminotransferase level): 75	ALP(Alkaline phosphatase level): 110	Iron: 90
Ferritin: 180		

Fig: 5 Prediction Page

This page collects user input for various parameters to predict the Viral Paradigm

RESULT PAGE:



Your New Value Prediction: Non Viral

Reasons for Prediction:

Category	Description
Cause	Alcohol Consumption, Obesity
Risk Factors	Cirrhosis, Diabetes
Symptoms	Unexplained Weight Loss, Jaundice, Abdominal Swelling
Diagnosis	Ultrasound, CT Scan, Liver Function Tests
Treatment	Lifestyle Changes, Weight Loss, Medications

Diet and Tips for Better

Tip	Description
1	Eat a balanced diet rich in fruits and vegetables
2	Incorporate lean protein sources like chicken and fish
3	Avoid processed foods and excessive sugar
4	Limit alcohol intake and maintain a healthy weight
5	Add fiber-rich foods like beans and legumes
6	Incorporate healthy fats such as avocado and nuts

Fig:6 Result Page

This page represent the output value of the input given by the user.

DOI: 10.48175/IJARSCT-25242

Confusion Matrix of Algorithms:

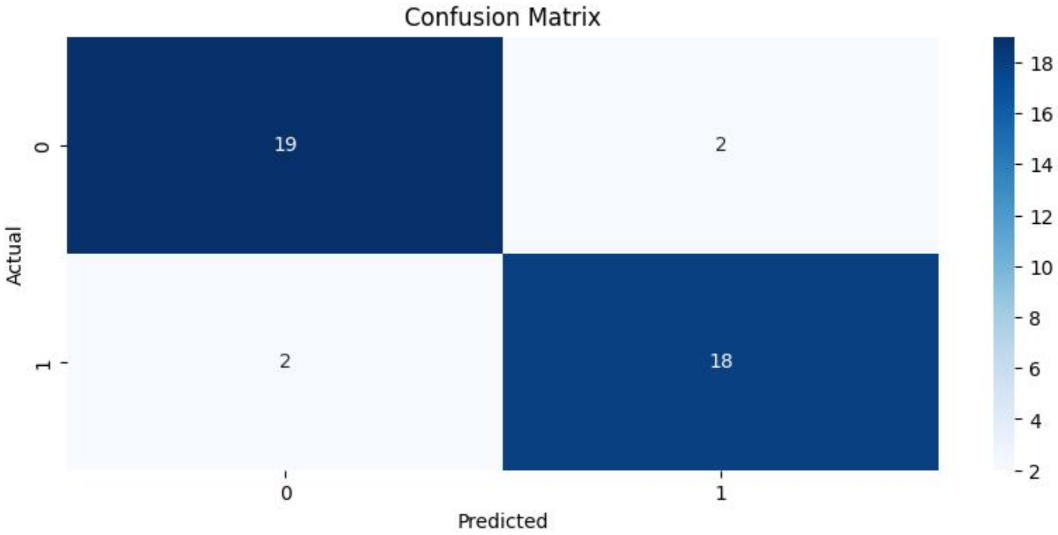


Fig:7 Confusion Matrix of Random Forest

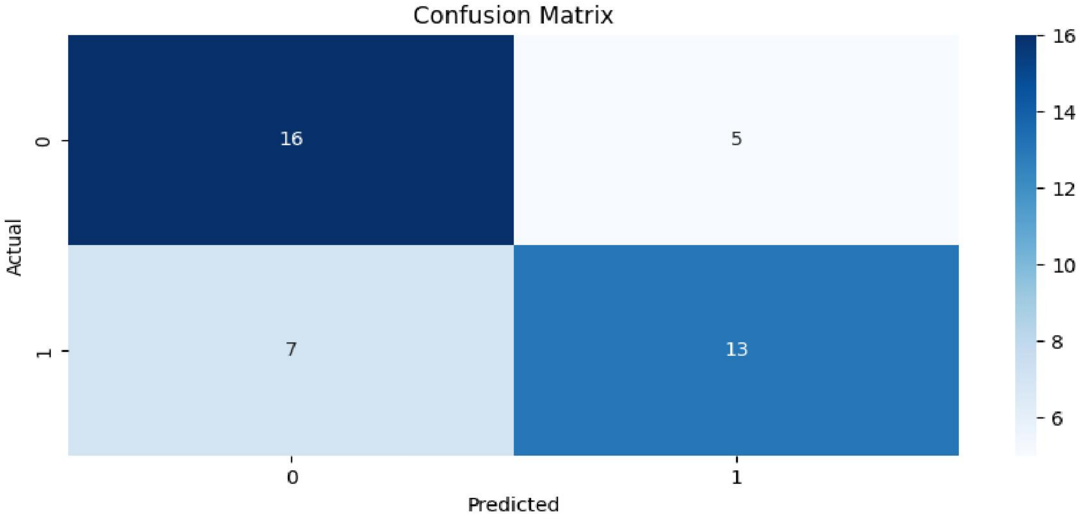


Fig:8 Confusion Matrix of Decision Tree

DOI: 10.48175/IJARSCT-25242

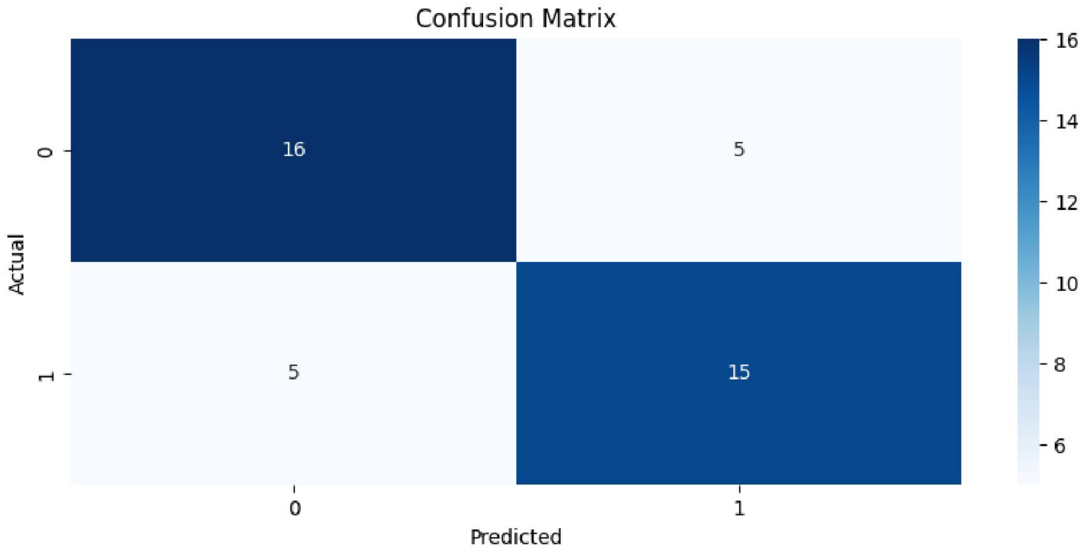


Fig:9 Confusion Matrix of Logistic Regression

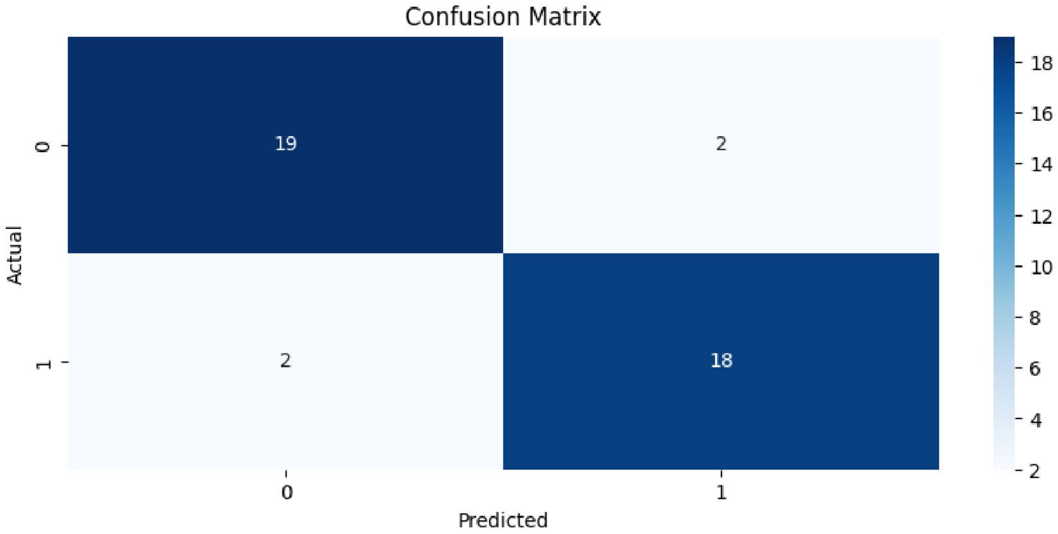


Fig:10 Confusion Matrix of Stacking Classifier

DOI: 10.48175/IJARSCT-25242

7.1.5 COMPARISON OF ALL MODELS

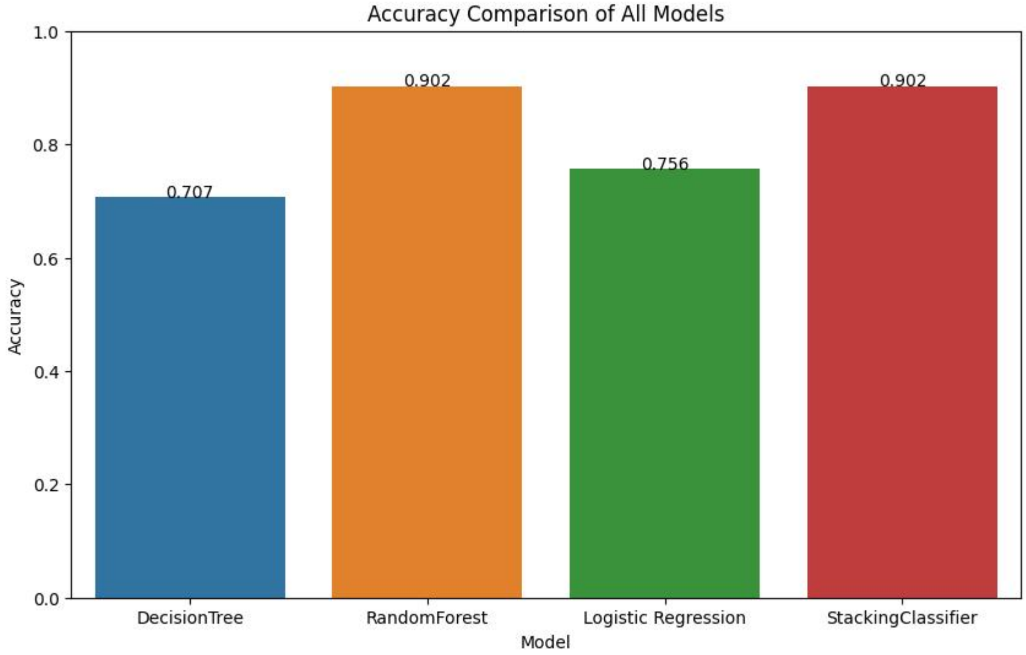


Fig:11 Accuracy of all models

The diagram presents an accuracy comparison of different machine learning 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.

Title: "Accuracy Comparison of All Models," highlighting model performance for predictive tasks.

The Random Forest and Stacking Classifier are the most effective based on accuracy.

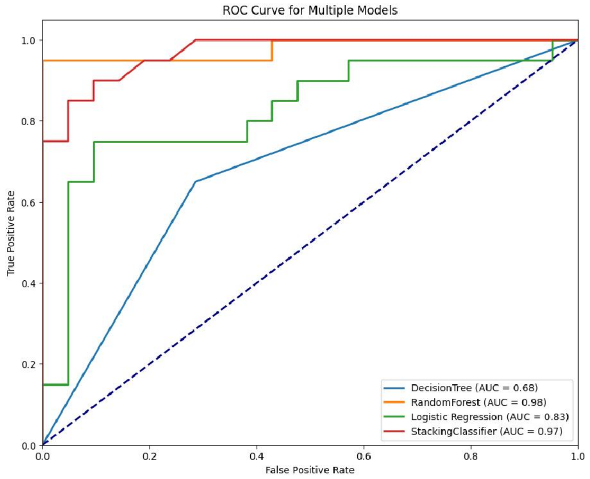


Fig:12 ROC and AUC curves of all models

The diagram illustrates the ROC (Receiver Operating Characteristic) curve for multiple classification models, comparing their performance in terms of True Positive Rate (TPR) against the False Positive Rate (FPR).
Axes: The x-axis represents the False Positive Rate, while the y-axis represents the True Positive Rate.
Curves: Different colored curves represent various models:
Decision Tree (Blue): AUC of 0.68, showing limited predictive performance.
Random Forest (Red): AUC of 0.98, indicating excellent classification ability.
Logistic Regression (Green): AUC of 0.83, showing good performance.
Stacking Classifier (Orange): AUC of 0.97, demonstrating high accuracy.
Dashed Line: Represents a baseline performance (random guessing), providing a point of reference for evaluating the models.

VI. CONCLUSION & FUTURE ENHANCEMENT

To wrap things up, creating a machine learning-based diagnostic system to tell apart viral and non-viral hepatocellular carcinoma (HCC) marks a big leap forward in oncology. Our study dives into a well-rounded dataset to assess how different classification algorithms—like Decision Tree, Random Forest, Logistic Regression, and a Stacking Classifier—perform. The findings show that machine learning can boost diagnostic accuracy beyond what traditional methods offer, giving doctors a powerful tool to distinguish between HCC types. This distinction is vital for crafting personalized treatment plans, ultimately enhancing patient care and outcomes. Our results highlight the transformative potential of machine learning in liver cancer diagnostics, setting the stage for more effective treatments and better survival rates. Looking ahead, we should aim to weave these models into clinical practices and investigate additional features that could further sharpen predictive abilities in HCC diagnosis.

As for future improvements to the machine learning diagnostic system for HCC, there are several promising directions to explore. First off, adding more features like genomic, proteomic, and clinical data could deepen our understanding of the disease and boost classification accuracy. We could also look into advanced algorithms, such as deep learning techniques and ensemble methods beyond just stacking, to further elevate predictive performance. Plus, using strategies like transfer learning could help us tap into pre-trained models on larger datasets, minimizing the need for a ton of labeled data.

Working closely with medical professionals for real-world validation and feedback will be crucial to fine-tune these models and ensure they're ready for clinical use.

Additionally, creating intuitive user interfaces for medical professionals might make it easier to incorporate this tool into clinical procedures. Last but not least, investigating the deployment of real-time monitoring systems and taking into account patient demographics and lifestyle characteristics may aid in tailoring treatment plans and enhancing patient outcomes in the management of HCC.

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.

DOI: 10.48175/IJARSCT-25242

[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.

[12] 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.

[13] 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

DOI: 10.48175/IJARSCT-25242