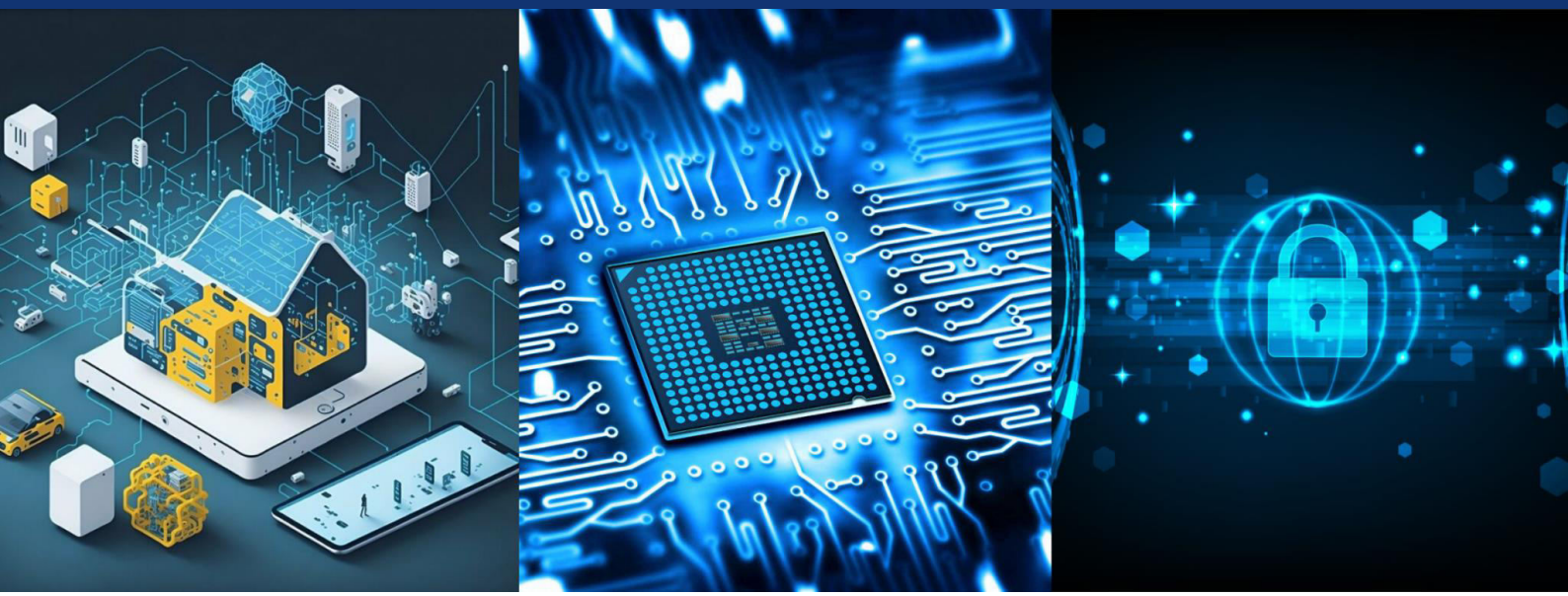
IJIRCCE

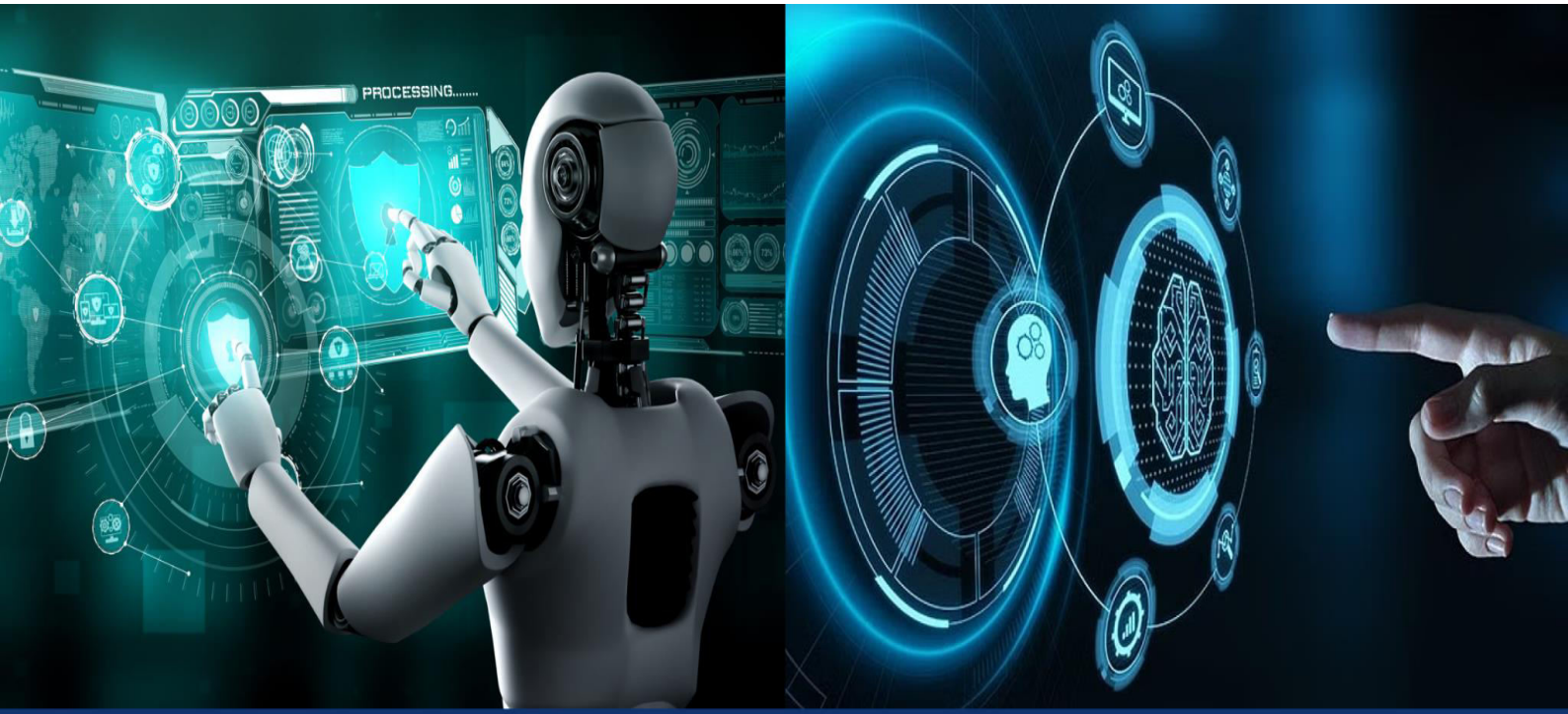
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 3, March 2025

© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303096

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Comparative Analysis of Diagnostic Techniques for Accurate Classification of Hepatocellular Carcinoma

K. Thrilochana Devi[1], K. Madhu Kondaiah[2], P. Amar Gopi[3], K.V.S. Praveen[4], K.Varun[5]

Assistant Professor, Department of IT, Vasireddy Venkatadri Institute of Technology, Guntur, India[1]

B.Tech. Student, Department of IT, Vasireddy Venkatadri Institute of Technology, Guntur, India[2,3,4,5]

ABSTRACT: Hepatocellular carcinoma (HCC) is a prevalent and aggressive form of liver cancer that arises from both viral factors, such as hepatitis infections, and non-viral factors, including fatty liver disease. Accurate diagnosis of HCC is vital for effective treatment and management; however, traditional diagnostic techniques like imaging and biopsies have inherent limitations. These methods often struggle with precision due to overlapping clinical and imaging features between viral and non-viral HCC, increasing the risk of misdiagnosis. To address this challenge, the project explores a machine learning-based approach to develop a diagnostic tool capable of accurately distinguishing between viral and non-viral HCC cases. Utilizing a balanced dataset from Kaggle, machine learning techniques like the Logistic Regression (LR), a random forest (RF), decision- tree (DT)and Stacking Classifier are employed to analyze and classify the data effectively. These models are trained and evaluated to ensure their accuracy in identifying HCC types, leveraging the ability of machine learning to process extensive datasets and deliver consistent results. Machine learning offers significant improvements over traditional diagnostic methods by enhancing the diagnostic process and reducing the likelihood of human error. This approach ensures greater precision and reliability, enabling personalized treatment strategies. The proposed system addresses the limitations of existing methods by improving diagnostic accuracy, streamlining workflows, and supporting informed clinical decision-making. By providing consistent and dependable results, this machine learning-driven diagnostic tool holds the potential to significantly improve patient outcomes and facilitate more effective treatment strategies for HCC.[1]

KEYWORDS: Hepatocellular carcinoma, Machine learning, Diagnostic tool, Decision Tree, Random Forest, Logistic Regression, Stacking Classifier, Diagnostic accuracy.

I. INTRODUCTION

One major global health concern is hepatocellular carcinoma (HCC) characterized by its varied etiology, including viral and non-viral origins, and its profound impact on liver function and patient prognosis. HCC severely affects individuals by compromising liver function, leading to complications such as jaundice, ascites, and hepatic encephalopathy, which ultimately result in a decline in quality of life. Early detection and intervention in HCC are crucial for improving patient outcomes. Timely diagnosis allows for rapid access to treatment and support services, enabling individuals and their families to better manage disease progression and optimize therapeutic strategies. Recent advancements in machine learning have revolutionized HCC diagnostics, offering new approaches to accurately classify the disease into viral and non-viral categories. By leveraging a comprehensive and balanced dataset, this project aims to apply cutting-edge machine learning Identify the best way to differentiate between these two forms of HCC. A thorough analysis of the most recent developments in machine learning-based methods for HCC detection and classification systematically evaluates the performance of various algorithms to identify the most effective methodologies. The algorithms used include Decision Tree (DT), which utilizes a tree-structured model to classify data and enable clear visualization of decision-making processes; Random Forest (RF), which, by averaging several predictions, uses an ensemble of decision trees to increase accuracy and decrease overfitting. Its capacity to manage big datasets and intricate relationships allows it to predict HCC categorisation with high accuracy.Logistic Regression (LR), which models binary outcomes using a logistic function and provides probabilistic interpretations of classification, being simple and interpretable but may not perform as well as more complex models in capturing non-linear relationships; and Stacking Classifier, which combines multiple algorithms to improve overall performance through meta-learning techniques and can leverage the strengths of various models but may require extensive computational resources and careful tuning. Our study aims to contribute to undertake continuous initiatives to enhance machine learning-based early HCC detection and categorisation. By assessing how well cutting-edge algorithms perform on actual datasets, we aim to provide a reliable, data-driven approach that supports

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

personalized treatment strategies. We further explore the potential implications of our findings for clinical practice, including the development of automated diagnostic tools and more targeted therapies. The project seeks to overcome current limitations in diagnostic accuracy, with the goal of lowering misclassification rates and enhancing patient care in general. Using the strength of sophisticated machine learning methods and thorough data analysis, we strive to revolutionize HCC diagnosis and treatment. Our ultimate goal is to provide healthcare professionals with the tools and insights needed to effectively combat this challenging liver cancer, contributing to better clinical outcomes and improved patient care.[2]

II. LITERATURE REVIEW

2.1 Related Work

Several studies have explored Ninety percent of liver tumours are hepatocellular carcinomas (HCCs), the most prevalent type of liver cancer and a major cause of death globally. Imaging is essential for monitoring and HCC diagnostic criteria, and early detection is critical. Various imaging modalities, including conventional ultrasound, multiphase CT, MRI, contrast-enhanced ultrasound (CEUS), CT and MR perfusion, elastography, T1 mapping, diffusion-weighted imaging (DWI), and MR spectroscopy, are essential for detecting and characterizing HCC. Emerging advanced imaging techniques like radiogenomics/radiomics aim to integrate quantitative radiologic data with clinical and immunobiological characteristics to improve prognosis and treatment outcomes. These advancements in imaging significantly contribute to the improved diagnosis and management of HCC, ultimately leading to better patient care[3].

2.2 Algorithmic Approaches

In the context of classifying Hepatocellular Carcinoma (HCC) into viral or non-viral categories, we employ models like Decision Tree, Random Forest, Logistic Regression, and Stacking Classifier. After evaluating these models, Random Forest was identified as having the highest classification performance. These models analyze preprocessed text data and user-provided inputs to distinguish between viral and non-viral HCC, enhancing diagnostic accuracy and informing personalized treatment strategies [4].

2.3 Datasets and Applications

To train robust machine learning models for the classification of Hepatocellular Carcinoma (HCC) into viral or non-viral categories, high-quality datasets are essential. We utilized a publicly available dataset containing various medical parameters for HCC research, which has been extensively used in related studies. However, the dataset's limited size and diversity present challenges for the generalization of the model. To address this issue, we explored data augmentation techniques to enhance the dataset, although further validation of these techniques is necessary[5].

2.4 Performance Metrics

Metrics including accuracy, precision, recall, and F1-score are typically used to assess performance when classifying hepatocellular carcinoma (HCC) into viral or non-viral groups.It is crucial to consider computational efficiency, especially for real-time applications, to ensure timely and accurate diagnostics. Emphasizing the need for robust evaluation frameworks is essential to guarantee reliable and reproducible results [6].

2.5 Research Gap

Despite significant progress, existing methods for classifying Hepatocellular Carcinoma (HCC) into viral or non-viral categories face several challenges, including low accuracy, high computational costs, and reliance on large, annotated datasets. This research aims to address these limitations by lev eraging models such as Decision Tree, Random Forest, Logistic Regression, and Stacking Classifier. Our approach is optimized for analyzing preprocessed text data and user-provided inputs, enhancing diagnostic accuracy and supporting personalized treatment strategies [7].

III. METHODOLOGY

The methodology for this project centers around developing a machine learning-based algorithm to classify Hepatocellular Carcinoma (HCC) into viral or non-viral categories. The approach involves utilizing various models, including Decision Tree, Random Forest, Logistic Regression, and Stacking Classifier. The methodology is divided into three key sections: Existing Algorithms, Proposed Algorithms, and Data Processing. Each section is thoroughly detailed, including relevant formulas, tables, and visualizations, to provide a comprehensive understanding of the approach.[8]

© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303096

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

3.1. Existing Algorithm

The existing algorithms for Hcc rely on imaging techniques, biopsy results, and clinical evaluations. These methods, while established, often face challenges in accurately differentiating between viral and non-viral HCC. Traditional systems may exhibit limitations such as subjective interpretation of imaging results, variability in biopsy outcomes, and the potential for misdiagnosis due to overlapping clinical features.[9]

The limitations of existing algorithms can be summarized as follows:

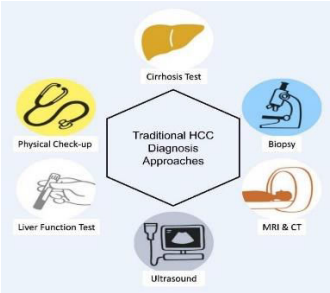
Subjective interpretation: Diagnostic results can be influenced by individual radiologists' and pathologists' interpretations.

Variability: Different diagnostic methods and clinicians may yield inconsistent results.

Misdiagnosis: There is a risk of incorrect classification of HCC types, leading to inappropriate treatment.

Limited precision: Traditional methods may lack the accuracy needed to reliably differentiate between viral and non-viral HCC.

Overlap: Clinical and imaging features of viral and non-viral HCC may be similar, complicating accurate diagnosis.



3.2. Proposed Algorithms

The proposed device goals to enhance HCC analysis by utilising machine learning algorithms to accurately classify hepatocellular carcinoma into viral and non-viral classes. This Utilising a balanced dataset from Kaggle, the system compares and assesses the efficacy of Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and a Stacking Classifier. Preprocessing the data, training several models, and evaluating each model's performance using metrics like accuracy, precision, recall, and F1-score are all part of the system. The very last output will be a strong category device that integrates the quality-performing version, presenting clinicians with a dependable and unique 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.[10]

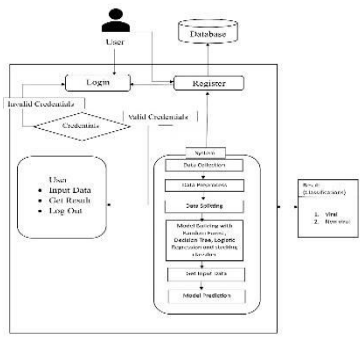


Figure 1:System Architecture of HCC Classification

© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303096

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

3.2.1 Random Forest Architecture

Random Forest is a type-classification and regression ensemble mastering order. In order to increase predicted accuracy and robustness, it builds several decision trees during training and aggregates their results.The algorithm follows the principle of bagging (Bootstrap Aggregating) to enhance model generalization and reduce overfitting.[11]

Construction of Random Forest:

Given a dataset The formula D={(X1,Y1),(X2,Y2),...,(Xn,Yn)} \{(X_1, Y_1), (X_2, Y_2),..., (X_n, Y_n)\} represents D. The random forest mode list constructed follows: D={(X1,Y1),(X2,Y2),...,(Xn,Yn)}, where Xi stands for the input features and YiY_iYi for the target labels.

Bootstrap-Sampling: Create m bootstrap samples D1,D2,...,DmD_1, D_2, ..., D_mD1,D2,...,Dm by randomly sampling DDD with replacement. Each sample is used to train an individual decision tree:

Decision Tree Learning: Each decision tree T_j (where j=1,2,...,m is trained independently on D_j At each node of a tree, a random subset of features is selected to determine the best split.

2. For classification: Classification: the final output is determined by the majority vote among all trees.

$$Y\hat{} = argkmax \sum_{j=1}^{m} 1\,(Tj(X) = k)$$[12]

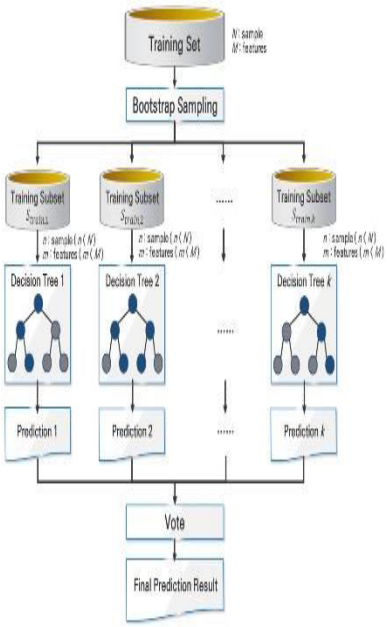


Figure.2 Random Forest Architecture

3.2.2 Decision Tree Architecture: A decision tree is an approach for supervised learning that may be applied to tasks involving both regression and classification. It creates a tree-like structure by dividing the dataset into smaller subsets according to feature conditions.[13]

Structure of a Decision Tree

1. **Root node:** denotes the starting point and the complete dataset.
2. **Internal Nodes**: Stand in for feature-based judgement.
3. Branches: Indicate possible outcomes of a decision.
4. Leaf Nodes: constitute final predictions (magnificence labels or numerical values).

Mathematical Formulation:

At each node, the best feature X_i is selected based on an impurity criterion.

© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303096

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

For Classification:

- **Gini Impurity**:

  Gini = 1-$\sum_{i=1}^{c} pi^2$

- **Entropy:**

  H= $-\sum_{i=1}^{c} pi \ \log_2 pi$

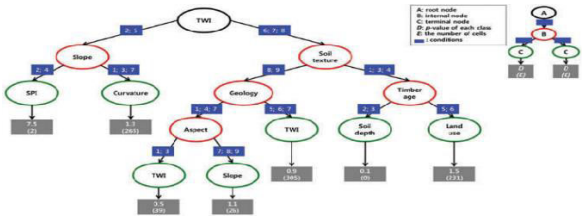


Figure.3 Decision Tree Architecture

3.2.3Logistic Regression: A statistical and knowledge-gathering approach for binary and multi-class category problems is called logistic regression. It calculates the likelihood that the internal nodes provide feature-based assessments that use the sigmoid function to determine if input belongs to a specific class.[14]

Mathematical Formulation

For a given input **X** with features $x_1, x_2, \cdots x_n$the model computes a weighted sum of the features:

$Z=w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n$

where:

- z is the linear sum of inputs xi and weights wi.
- The bias term (intercept) is w0.

An output between 0 and 1 is obtained by applying the logistic function (sigmoid) to z:

$\sigma(z)=1/1 + e^{-z}$

where σ(z) is the probability of the positive class P(Y=1|X)

For **classifications**, we set a decision boundary:

- Classify it as 1 (positive class) if σ(z)≥0.5
  and as 0 (negative class) if σ(z)<0.5.

3.2.4 Stacking Classifier: stacking classifier is an ensemble learning technique that improves predictive accuracy by combining a few basic models. In contrast to boosting (like X-GBoost) or bagging (like Random Forest), stacking uses a meta-learner, also known as a blender, to learn how to combine multiple models.

Mathematical Representation

Let **X** be the input features and **Y** be the target labels.

1. **Base models** $f_1, f_2, \cdots f_3$ generate predictions:

   $$P_1 = f_1(X), P_2 = f_2(X), P_n = f_n(X)$$

2. These predictions become the new feature set for the meta-learner g:

   $$Y_{final} = g(P_1, P_2, \cdots P_N)$$

where g is the **meta-classifier**, which learns how to best combine predictions.

© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303096

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

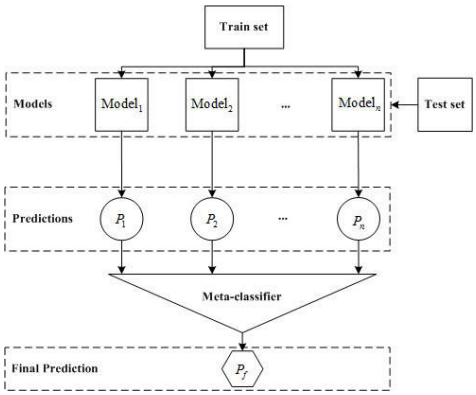


Figure.4 Stacking Classifier Architecture

3.3. Mixed Precision Training

To enhance the efficiency and performance of machine mastering algorithms together with Random forest, decision Tree, Logistic Regression, and Stacking Classifier, mixed precision training can be applied. This approach leverages 16-bit floating-point numbers (float16) for specific computations to reduce memory usage and speed up training while maintaining accuracy. It is particularly effective when optimizing hyperparameters, allowing for faster tuning and improved scalability, especially on GPUs.[15]

The following formula can be used to summarise the mixed precision training process:

$\hat{y} = f(x; \theta\_16)$ εq (1)

$\nabla\_(\theta\_16)\, L = \partial L/(\partial \theta\_16)$ εq (2)

$\theta\_32 = \theta\_32 - \eta \nabla \theta\_16$ εq (3)

Where:

- $\hat{y}$ is the predicted output of the algorithm.
- $f(x; \theta\_16)$ represents the forward pass with 16-bit precision for computation.
- $\nabla\_(\theta\_16)\, L$ is the loss function's gradient in relation to the 16-bit weights..
- θ_32 represents the 32-bit weights used for the weight update.
- n is the learning rate

3.4. Model Evaluation

To evaluate the algorithm's overall performance, a test dataset with scientific inputs in.csv format is used. The evaluation measures, such as accuracy, precision, recall, and the F1 score, are calculated using the following mathematical formulas:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$

Where:

- TP stands for True Positives
- TN for True Negatives
- FP for False Positives
- FN for False Negatives

The classification file and confusion matrix are also produced in order to provide a thorough analysis of the model's overall performance.

3.5. Data Processing

3.5.1. Dataset Description

The HCC dataset consists of 204 samples with 50 functions, categorized into 34 numerical and sixteen categorical variables. The target variable, "Class", shows the presence of Hepatocellular Carcinoma (HCC).The dataset includes

© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303096

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

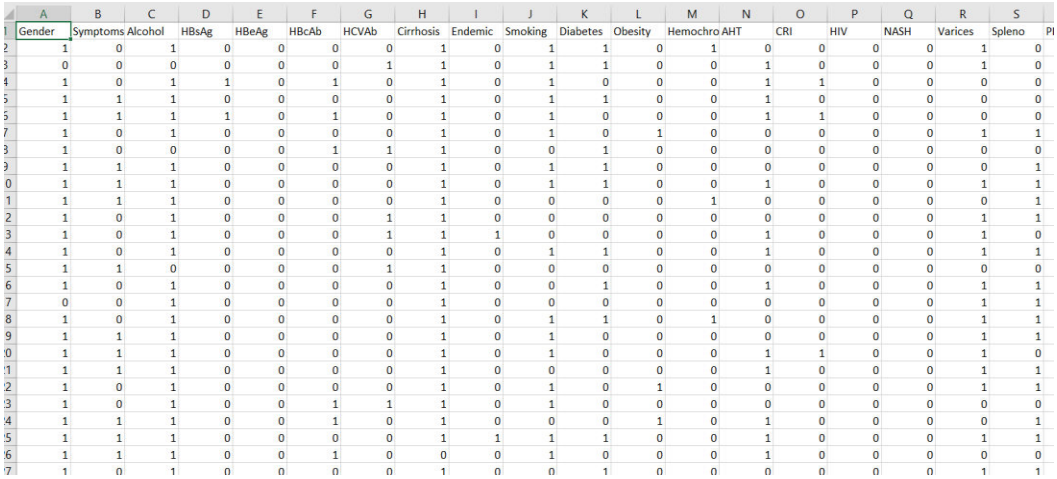
International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

medical attributes such as age, liver function markers (ALT, AST, ALP), viral infection markers (HBsAg, HCVAb), and lifestyle factors (smoking, alcohol consumption). Some numerical values are stored as strings with commas instead of decimal points, requiring preprocessing. Additionally, the dataset is imbalanced, meaning the number of samples in each class varies, which may impact model performance. To ensure effective training and evaluation, the dataset needs cleaning, feature transformation, and handling of imbalances.[16]

Table 3.1 Dataset

Dataset	Viral HCC Cases	Non-Viral HCC Cases	Total Cases
Training	81	81	162
Validation	10	10	20
Test	11	11	22



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Gender	Symptoms	Alcohol	HBsAg	HBeAg	HBcAb	HCVAb	Cirrhosis	Endemic	Smoking	Diabetes	Obesity	Hemochro	AHT	CRI	HIV	NASH	Varices	Spleno
2	1	0	1	0	0	0	0	1	0	1	1	0	1	0	0	0	0	1	0
3	0	0	0	0	0	0	1	1	0	1	1	0	0	1	0	0	0	1	0
4	1	0	1	1	0	1	0	1	0	1	0	0	0	1	1	0	0	0	0
5	1	1	1	0	0	0	0	1	0	1	1	0	0	1	0	0	0	0	0
6	1	1	1	1	0	1	0	1	0	1	0	0	0	1	1	0	0	0	0
7	1	0	1	0	0	0	0	1	0	1	0	1	0	0	0	0	0	1	1
8	1	0	0	0	0	1	1	1	0	0	1	0	0	0	0	0	0	0	0
9	1	1	1	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	1
10	1	1	1	0	0	0	0	1	0	1	1	0	0	1	0	0	0	1	1
11	1	1	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1
12	1	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1
13	1	0	1	0	0	0	1	1	1	0	0	0	0	1	0	0	0	1	0
14	1	0	1	0	0	0	0	1	0	1	1	0	0	1	0	0	0	1	1
15	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
16	1	0	1	0	0	0	0	1	0	0	1	0	0	1	0	0	0	1	1
17	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1
18	1	0	1	0	0	0	0	1	0	1	1	0	1	0	0	0	0	1	1
19	1	1	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	1
20	1	1	1	0	0	0	0	1	0	1	0	0	0	1	1	0	0	1	0
21	1	1	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	1
22	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1
23	1	0	1	0	0	1	1	1	0	1	0	0	0	0	0	0	0	0	0
24	1	1	1	0	0	1	0	1	0	0	0	1	0	1	0	0	0	0	1
25	1	1	1	0	0	0	0	1	1	1	1	0	0	1	0	0	0	1	1
26	1	1	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0
27	1	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	1

Figure.5 Sample image of Training

The dataset sample images used for implementing the algorithms.

	Gender	Smoking	AFP	Hemoglob	MCV	Albumin	AST	ALP	Iron	Ferritin
158	1	0	1713	8.2	94.2	3.6	59	263	9	490
49	1	1	20	15	96.7	4.6	49	109	184	905
34	1	0	5689	14.3	99.6	3.8	102	184	94	48
154	0	0	615	11.7	99.7	2.82	50	318	25	60
111	1	1	173	11.1	105.5	3.1	206	188	105	221
8	1	1	8.8	11.9	107.5	1.9	59	63	85	982
113	1	1	42	16.2	99.1	4.1	118	158	40	57
189	1	1	60.9	11.9	98.9	2.44	66	239	118	748
156	1	0	975	15.3	103	3.5	85	266	180	1176
60	1	1	608	12.6	100	4	99	100	161	297
132	0	0	152	10.9	99.6	3.2	80	106	52.5	856
98	1	0	5.5	13.1	89.6	3.2	30	92	93	29
115	1	1	736	14	88.2	4.7	113	629	184	905
0	1	1	95	13.7	106.6	3.4	41	150	52.5	856
56	1	1	2.7	7.3	90.8	3.4	32	55	22	48
74	1	1	2.5	14.9	92.3	3.8	38	101	61	255
17	1	1	9.2	10.3	103.7	3.8	91	146	187	443
125	1	1	421500	14.3	89.5	3.1	44	217	52	832
159	1	0	4.9	7.9	111.2	2.43	71	73	40	283
70	0	0	358	12.7	74	2.9	41	94	50	20
190	1	0	2.8	11.8	92.3	3.44	80	307	77.5	864
198	1	0	337.2	12.4	90.6	3.45	228	575	137.9	718
64	1	0	7	12.1	95.1	3.8	38	161	0	0
53	1	0	7.3	15.3	93.7	2.1	93	130	184	59
161	0	0	4887	12.1	88.9	3	91	280	94	48
194	1	0	283	8.4	112.2	2.51	106	85	56.9	366

Figure. 6 Sample image of Validation

© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303096

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

	Gender	Smoking	AFP	Hemoglob	MCV	Albumin	AST	ALP	Iron	Ferritin
158	1	0	1713	8.2	94.2	3.6	59	263	9	490
49	1	1	20	15	96.7	4.6	49	109	184	905
34	1	0	5689	14.3	99.6	3.8	102	184	94	48
154	0	0	615	11.7	99.7	2.82	50	318	25	60
111	1	1	173	11.1	105.5	3.1	206	188	105	221
8	1	1	8.8	11.9	107.5	1.9	59	63	85	982
113	1	1	42	16.2	99.1	4.1	118	158	40	57
189	1	1	60.9	11.9	98.9	2.44	66	239	118	748
156	1	0	975	15.3	103	3.5	85	266	180	1176
60	1	1	608	12.6	100	4	99	100	161	297
132	0	0	152	10.9	99.6	3.2	80	106	52.5	856
98	1	0	5.5	13.1	89.6	3.2	80	92	93	29
115	1	1	736	14	88.2	4.7	113	629	184	905
0	1	1	95	13.7	106.6	3.4	41	150	52.5	856
56	1	1	2.7	7.3	90.8	3.4	32	55	22	48
74	1	1	2.5	14.9	92.3	3.8	38	101	61	255
17	1	1	9.2	10.3	103.7	3.8	91	146	187	443
125	1	1	421500	14.3	89.5	3.1	44	217	52	832
159	1	0	4.9	7.9	111.2	2.43	71	73	40	283
70	0	0	358	12.7	74	2.9	41	94	50	20
190	1	0	2.8	11.8	92.3	3.44	80	307	77.5	864
198	1	0	337.2	12.4	90.6	3.45	228	575	137.9	718
64	1	0	7	12.1	95.1	3.8	38	161	0	0
53	1	0	7.3	15.3	93.7	2.1	93	130	184	59
161	0	0	4887	12.1	88.9	3	91	280	94	48
194	1	0	283	8.4	113.2	2.51	106	85	56.9	366

Figure.7 Sample image of Testing

3.5.2. Data Augmentation and Preprocessing

To enhance the version's performance on the HCC dataset, numerous information preprocessing techniques are carried out to beautify generalization and reduce overfitting. These techniques ensure that the model learns robust patterns, even with limited and imbalanced data. Since the dataset consists of numerical and categorical medical records rather than images, preprocessing involves data normalization, feature scaling, and augmentation strategies like synthetic data generation.

The preprocessing pipeline can be summarized as follows:

1. **Handle Missing Values:** Identify and fill missing values using median imputation for numerical features and mode imputation for categorical features.
2. **Convert Categorical Data:** Encode express variables the usage of one-warm encoding or label encoding as needed.
3. **Normalize Numerical Features:** Scale continuous numerical features to a standard range (0,1) using Min-Max Scaling or Standardization for better convergence during training.
4. **Handle Class Imbalance:** To ensure balanced model learning, create synthetic samples for under-represented classes using the Synthetic Minority Over-sampling Technique (SMOTE).
5. **Feature Selection:** Remove irrelevant or highly correlated features to reduce dimensionality and improve computational efficiency.[17]

3.5.3. Data Loading and Batching

To guarantee effective training and ideal model performance, the HCC dataset—which is saved in a.csv file—is loaded and pre-processed. The dataset is first read into a pandas Da-taFrame, where the mode is used to impute values into categorical columns and the median is used to fill in missing values in numerical columns. In order to ensure interoperability with device learning models, categorical features, such gender, are encoded using Label Encoding or One-hot Encoding. Min-Max Scaling is used to normalise numerical features to a 0–1 range in order to increase training stability.

Stratified sampling is then used to preserve class distribution while the dataset is divided into training (80%), validation (10%), and test (10%) sets. The statistics are similarly batched into agencies of 32 samples to optimise computational performance, lowering reminiscence load and increasing processing velocity. When training machine learning models like Random Forest, Decision Tree, Logistic Regression, and Stacking Classifier, this systematic pipeline guarantees that the dataset is clear, well-structured, and processed effectively.[18]

© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303096

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. OUTCOMES AND DEFINITIONS

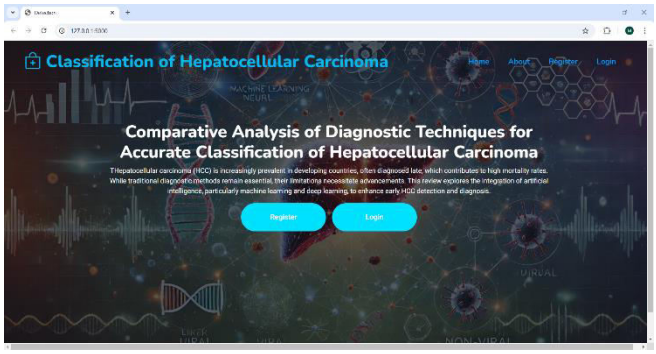


Figure.8 Index Page

The following image output shows the register page where user has to give basic information

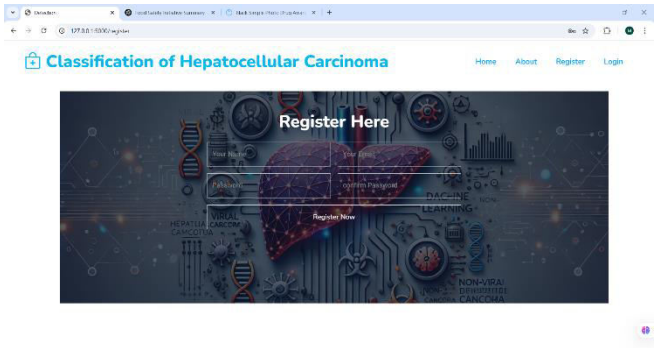


Figure.9 Register Page

The below page shows the dataset values used for training and testing the algorithms.

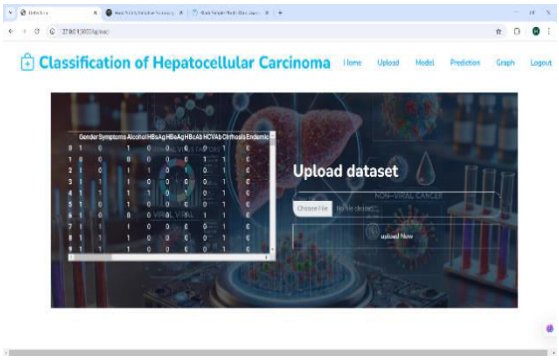


Figure.10 Upload Dataset

This page indicates what inputs has to given by user to predict the HCC is Viral or Non-Viral.

© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303096

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Figure 11. Prediction page

This page provide Results from collecting information for predicted page.

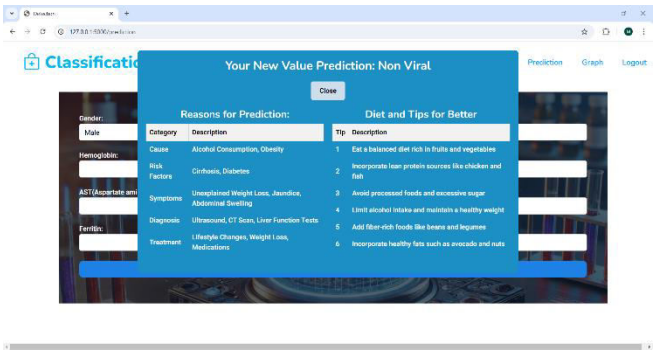


Figure 12. Results

4.1 Output Graphs and Visualizations

To monitor the model's performance, accuracy and loss curves are plotted for both training and validation phases. Additionally, a confusion matrices, ROC and AUC curves and classification report are generated to evaluate its effectiveness on the test dataset.

- **ROC and AUC Analysis**: The ROC curves for multiple models are plotted to compare their performance in distinguishing between classes. The AUC (Area Under the Curve) scores indicate the effectiveness of each model, with higher values representing better classification capability. The Stacking Classifier achieves the highest AUC (0.97), followed by Random Forest (0.98), Decision Tree (0.71) and Logistic Regression (0.83). The best model for classification is chosen with the aid of this analysis.[19]

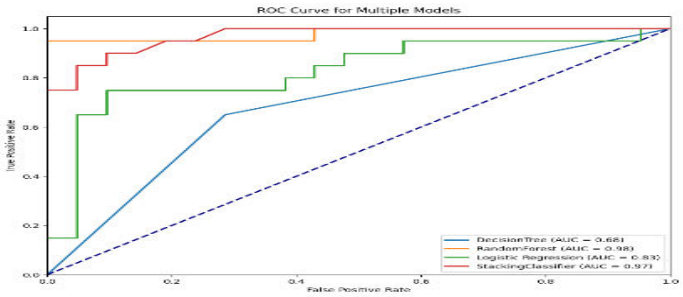


Figure13. ROC and AUC Plot

© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303096

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The below Bar-Graph represents the Accuracy Comparision of All Model's.

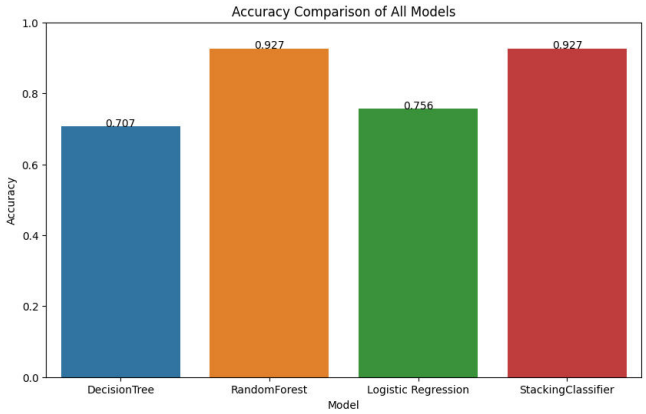
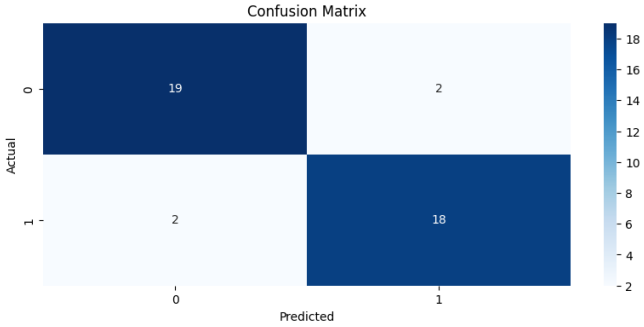
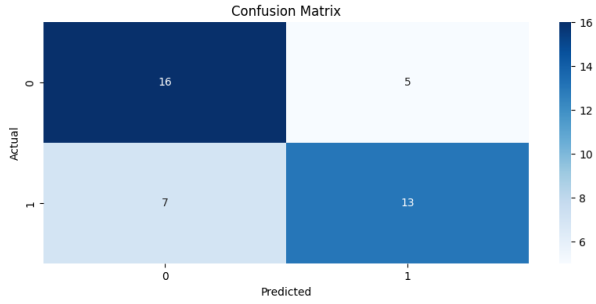


Figure 14 Accuracy

- Confusion Matrices: displays the proportion of accurate and inaccurate predictions for every class.



The Above Figure Show's the Confusion Matrix of Random Forest.



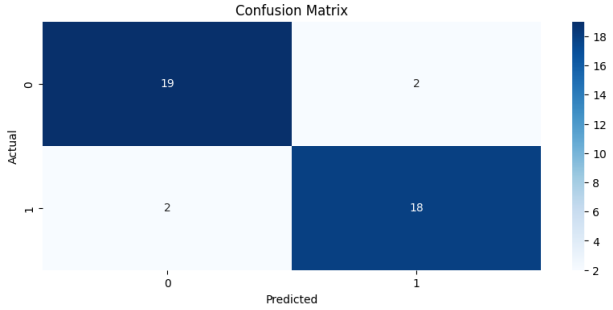
The Above Figure Show's the Confustion Matrix of Decision Tree.

© 2025 IJIRCCE | Volume 13, Issue 3, March 2025| DOI: 10.15680/IJIRCCE.2025.1303096

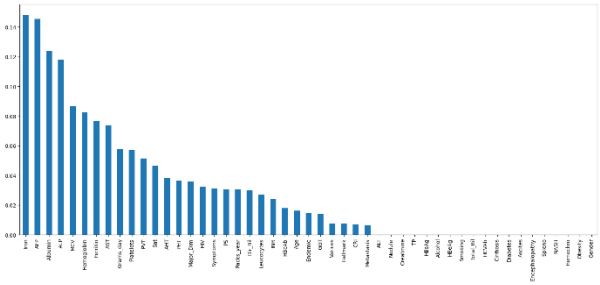
www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



The Above Figure Show's the Confustion Matrix of Stacking Clasifier.



The above Figure Shows the Feature Selection, where it used as input by the User to Predict the New Value.

- Model's Comparison For Classification:

Table 4.1 Classification Report:

Model	Precision	Recall	F1-Score	Accuracy
Random Forest	0.99	0.90	0.91	0.92
Logistic Regres-sion	0.75	0.75	0.75	0.75
Decision Tree	0.76	0.65	0.70	0.70
Stacking Classifier	0.90	0.85	0.87	0.90

The proposed algorithm leverages the Random Forest architecture, showcasing its superior predictive capability through ensemble learning, which reduces overfitting and enhances accuracy. Designed to process structured input data, it is highly effective for analyzing medical and clinical datasets. By incorporating automated feature selection and hyperparameter tuning, the algorithm eliminates the need for manual preprocessing, optimizing the classification process. Additionally, a Flask-based web application is implemented, allowing users to input data and receive real-time predictions using the Random Forest model. This approach ensures a scalable, efficient, and reliable solution for medical data analysis, leading to improved decision-making and predictive accuracy.[20]

V. CONCLUSION

In summary, a major breakthrough in oncology has been made with the creation of a machine learning-based diagnostic paradigm for differentiating between viral and non-viral hepatocellular carcinoma (HCC). Our work methodically assesses the performance of several classification algorithms, such as Decision Tree, Random Forest, Logistic Regression, and a Stacking Classifier, by utilising a comprehensive and balanced dataset. The results demonstrate that machine learning techniques can enhance diagnostic accuracy beyond traditional methods, offering clinicians a robust tool

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

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.[21]

VI. FUTURE SCOPE

Future enhancements of the machine learning-based diagnostic paradigm for hepatocellular carcinoma (HCC) could focus on several key areas. Firstly, integrating additional features such as genomic, proteomic ,Incorporating clinical data could enhance classification accuracy and offer a more thorough knowledge of the illness. Implementing advanced algorithms, such as deep learning techniques and ensemble methods beyond stacking, could further enhance predictive performance. Additionally, employing techniques like transfer learning could leverage pre-trained models on larger datasets, reducing the need for extensive labeled data.Collaboration with medical professionals for real-world validation and feedback will be essential to refine the models and ensure clinical applicability. Moreover, developing user-friendly interfaces for healthcare practitioners can facilitate the integration of this tool into clinical workflows. Lastly, exploring the implementation of real-time monitoring systems and incorporating patient demographics and lifestyle factors could help in personalizing treatment strategies and improving patient outcomes in HCC management**.[22]**

REFERENCES

[1] HCC base paper with the aid of Criss et al. (2023) specializes in the diagnostic imaging of Hepatocellular Carcinoma (HCC) using numerous radiological strategies which include MRI, CT, and ultrasound. The paper dis-cusses the significance of earlydetection in im-proving affected person consequences and explores superior imaging strategies like radio genomics, perfusion imaging, and comparison-greater ultrasound to ehance the accuracy of HCC prognosis.

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

[3] J. D. Yang, P. Hainaut, G. J. Gores, A. Amadou, A. Plymoth, and L. R. Roberts, ‘‘A international view ofhepatocellular carcinoma: tendencies, risk, prevention and control,’’ Nature Rev. Gastroenterol. Hepatol., vol. 16,no. 10, pp. 589–604, Oct. 2019.

[4] [4] I. Sghaier, S. Zidi, L. Mouelhi, E. Ghazoueni, E. Brochot, W. Almawi, and B. Loueslati, ‘‘TLR3 and TLR4 SNP versions inside the liver disease as a result of hepatitis B virus and hepatitis C virus infec-tion,’’ Brit. J. Biomed. Sci.,vol. 76, no. 1, pp. 35–forty one, Jan. 2019.

[5] "Purinoceptor expression in hepatocellular virus (HCV)-precipitated and non-HCV-precipitated hepatocellular carcinoma: An insight into the proviral position of the P2X4 receptor," by M. Khalid, S. Manzoor, H. Ahmad, A. Asif, T. A. Bangash, A. Latif, and S. Jaleel December 2018, Mol. Biol. Rep., vol. 45, no. 6, pp. 2625–2630.

[6] A. Asif, M. Khalid, S. Manzoor, H. Ahmad, and A. U. Rehman, ‘‘role of purinergic receptors in hepatobiliarycarcinoma in Pakistani population: An method toward proinflammatory function of P2X4 and P2X7 receptors,’’Purinergic Signalling, vol. 15, no. three, pp. 367–374, Sep. 2019.

[7] 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..

[8] Hamesch and P. Strnad, ‘‘Non-invasive examine-ment and control of liver involvement in adults withAlpha-1 antitrypsin deficiency,’’ continual Obstructive Pulmonary sicknesses: J. COPD discovered., vol. 7, no. 3, pp. 260–271, 2020.

[9] Patel and G. Sebastiani, ‘‘barriers of non-invasive assessments for evaluation of liver fibrosis,’’ JHEP Rep., vol. 2,no. 2, Apr. 2020, artwork. no. 100067.

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

[11] Y. Masugi, T. Abe, H. Tsujikawa, okay. 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, ‘‘Quantita-tive precise 48% Plagiarized 52% assessment of liver fibrosis exhibits a nonlinear affiliation with fibrosis degree in nonalcoholic fats-ty liver sickness,’’ Hepatology Commun., vol. 2, no. 1, pp. 58–68, 2018. [11] k. Y. Ngiam and i. W. Khor, ‘‘big rec-

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

ords and system mastering algorithms for fitness-care shipping,'' Lancet Oncol., vol. 20, no. five, pp. e262–e273, can also 2019..

[12] M. Subramanian, A. Wojtusciszyn, L. Favre, S. Boughorbel, J. Shan, ok. B. Letaief, N. Pitteloud, and L. Chouchane, ''Precision medicine within the era of synthetic intelligence: Implications in chronic dis-ease management,'' J. Transl. Med., vol. 18, no. 1, pp. 1–12, Dec. 2020.

[13] H. B. El-Serag, J. A. Marrero, L. Rudolph, and okay. R. Reddy, ''analysis and remedy of hepatocellular carcinoma,'' Gastroenterology, vol. 134, no. 6, pp. 1752–1763, 2008.

[14] writer T, et al. "scientific implications of osteopo-rosis detection the usage of deep studying." The Lancet digital fitness,2021.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details