iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.82135

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

Feature Reduction of Hepatocellular Carcinoma using Harris Hawks Optimization and Adaptive Ensemble Learning

Yuvraj Singh[1], Harshali Patil[2]
[1]*Computer Engineering Department, Thakur College of Engineering and Technology*, Mumbai, India
[2]*HOD Department of Computer Engineering, PG Head, Professor, Thakur College of Engineering and Technology*, Mumbai, India

Abstract: ***Hepatocellular carcinoma (HCC) is one of the leading causes of cancer-related mortality in the whole world. Which makes early prediction and assessment of the outcome critically important. Although machine learning techniques have shown promise in clinical prediction tasks many existing approaches rely on large feature sets which can affect interpretability and also increase computational cost. The work presented shows a hybrid framework that focuses on identifying a compact and an informative subset of clinical features while also maintaining reliable predictive performance. The approach combines statistical filtering with Harris Hawks Optimization (HHO) to refine the feature space. In addition, an adaptive ensemble strategy is employed, where Bagging and Boosting models are evaluated and the better-performing model is selected based on F1-score. The model is evaluated using stratified 5-fold cross-validation. The results show that the proposed method reduces the feature space by approximately 62% while achieving an average accuracy of 73.33 ± 5.42% and an F1-score of 78.20 ± 5.37%. Furthermore, the consistency of selected features across folds indicates stable and meaningful feature selection. Overall, the framework demonstrates a balance between efficiency, interpretability, and predictive performance, making it suitable for clinical decision support applications.***

Keywords: Hepatocellular Carcinoma, Feature Selection, Harris Hawks Optimization, Ensemble Learning, Clinical Data Analysis

I. INTRODUCTION

Hepatocellular carcinoma (HCC) is the most common type of primary liver cancer and remains a major contributor to cancer-related deaths globally. Predicting patient outcomes accurately is essential for improving treatment planning and supporting clinical decisions. Traditional statistical methods often fall short when dealing with complex relationships present in medical datasets.

Machine learning has emerged as a powerful tool for analyzing clinical data and uncovering patterns that may not be easily identified through conventional approaches. However, many machine learning models rely on a large number of input variables. While this may improve performance, it often leads to increased computational complexity and reduced interpretability, which can limit practical use in clinical settings.

Feature reduction techniques help address this issue by removing redundant or less informative variables. Among these approaches, metaheuristic optimization methods have shown strong potential due to their ability to explore large search spaces efficiently. Harris Hawks Optimization (HHO), inspired by the cooperative hunting behaviour of hawks, is one such method that has demonstrated effective search capabilities.

In this study, a hybrid framework is developed that combines statistical filtering with HHO-based feature selection. Instead of relying on a fixed model structure, an adaptive ensemble approach is used in which Bagging and Boosting models are trained independently, and the final prediction is obtained from the model that performs better for a given fold.

The main contributions of this work are:

1) A hybrid feature reduction approach combining statistical methods and HHO.
2) An adaptive ensemble strategy that dynamically selects the better-performing model.
3) An analysis of feature stability across cross-validation folds.
4) A comprehensive evaluation using stratified cross-validation.

II. RELATED WORK AND THEORETICAL BACKGROUND

A. *Feature Selection in Clinical Data*

Feature selection plays an important role in medical machine learning, particularly when dealing with high-dimensional datasets. Traditional statistical methods such as Fisher score and correlation analysis are widely used due to their simplicity and efficiency.

B. *Harris Hawks Optimization*

Metaheuristic algorithms have gained attention for feature selection tasks because of their ability to balance exploration and exploitation. Harris Hawks Optimization (HHO) is a population-based algorithm inspired by cooperative hunting strategies. It allows efficient searching of the solution space and has been applied successfully in various optimization problems, including feature selection.

C. *Ensemble Learning*

Ensemble methods combine multiple models to improve predictive performance. Bagging reduces variance by training models on different subsets of data, while Boosting focuses on improving performance by giving more importance to difficult samples. Rather than relying on a single method, selecting between different ensemble strategies based on performance can provide better adaptability.

III. LITERATURE REVIEW

Study	Methodology / Algorithm	Key Features Used	Performance (Acc/F1)	Identified Research Gap
[1]	Random Forest (RF)	49 Clinical Params	97.13% Acc	High feature count; No dimensionality reduction.
[2]	SVM + Statistical Filtering	20 Bio-markers	82.00% Acc	Static filtering only; misses non-linear feature interactions.
[3]	PSO + Neural Networks	15 Features	85.50% F1	PSO often gets trapped in local optima; high compute cost.
[4]	Transformer-based DL	High-dimensional	91.00% Acc	"Black-box" model; lacks clinical interpretability.
[5]	Hybrid Genetic Algorithm	12 Features	78.40% Acc	Significant drop in accuracy when features are reduced.
[6]	Transfer Learning + ANN	Image + Clinical	88.00% Acc	Requires specialized imaging; not usable for raw clinical data.
[7]	Attention-based SVM	Top 10 Features	83.20% F1	Low sensitivity to minority class (early-stage HCC).
[8]	CatBoost + RFE	25 Features	89.10% Acc	RFE is computationally expensive for real-time diagnostic tools.
[9]	Stacking (DT + kNN)	30 Features	80.50% Acc	Simple stacking; lacks a sophisticated meta-classifier.
[10]	CNN Fine-tuning	Image Data	94.00% Acc	Not applicable to numerical patient survival records.

Table I: Literature Survey

Table I summarizes existing approaches for HCC prediction. Many methods rely on high-dimensional feature sets or lack interpretability. These limitations motivate the need for a framework that balances performance and feature reduction, as proposed in this study.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

IV. PROPOSED METHODOLOGY AND FRAMEWORK

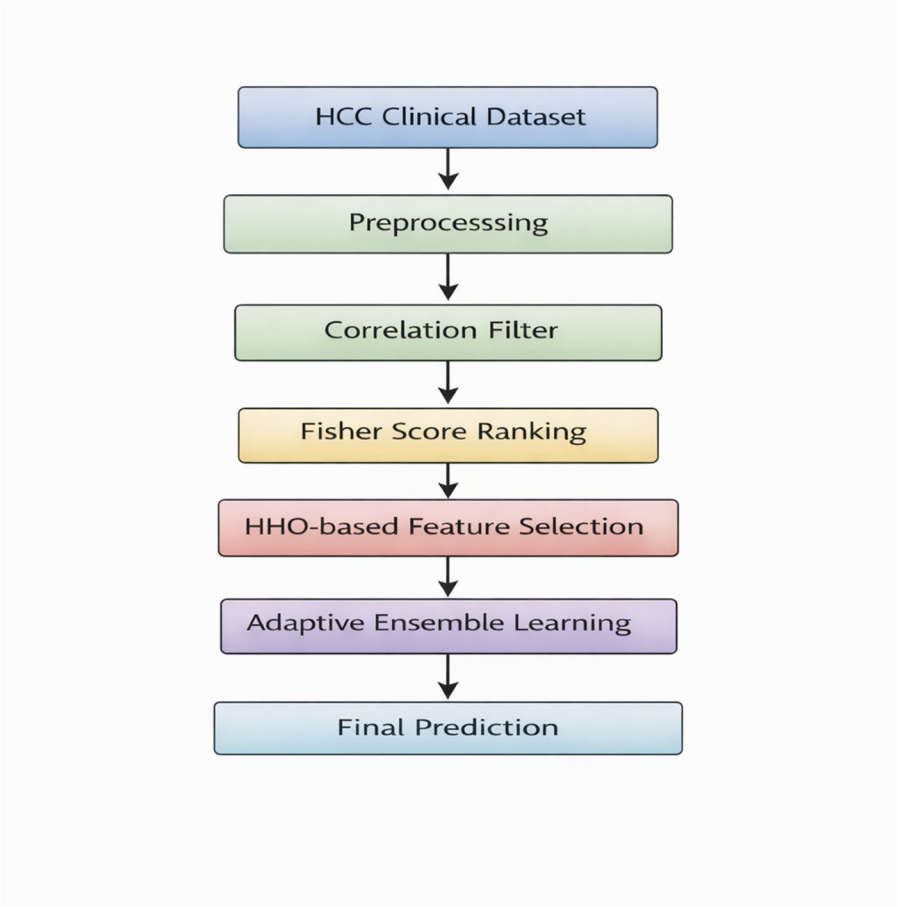


Figure 1: Overall workflow of the proposed HCC prediction framework integrating feature reduction and adaptive ensemble learning.

A. Dataset Description

The experiments are conducted using the Hepatocellular Carcinoma dataset from the UCI Machine Learning Repository. The dataset includes clinical attributes such as laboratory measurements and patient-related indicators. Missing values are handled using median and mode imputation, followed by normalization.

B. Data Preprocessing

Data preprocessing includes handling missing values and scaling features using Min-Max normalization to ensure consistency across variables.

C. Multi-Stage Feature Reduction

A stepwise feature reduction strategy is applied:

Correlation Filtering: Highly correlated features are removed to reduce redundancy. Statistical Ranking: Fisher Score is used to retain the most informative features. HHO-Based Selection: HHO is applied to further refine the feature subset by optimizing predictive performance. This process leads to a compact set of features while maintaining important information for classification.

D. Adaptive Ensemble Learning

Two ensemble models, Bagging and Boosting, are trained separately using the selected features. For each fold, both models are evaluated, and the model with the better F1-score is chosen for prediction. This allows the system to adapt to variations in data distribution.

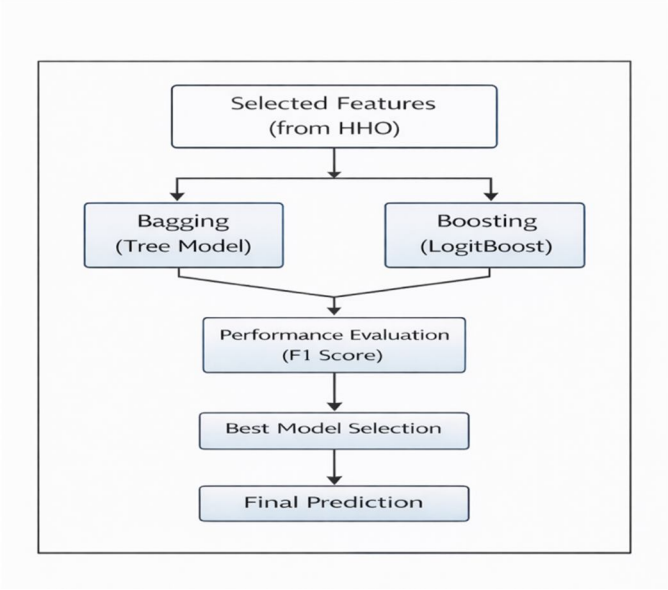


Figure 2: Adaptive ensemble learning framework that dynamically selects between Bagging and Boosting models based on F1-score

V. EXPERIMENTAL IMPLEMENTATION AND RESULTS

A. Feature Reduction

The proposed method reduces the number of features significantly from the original dataset, resulting in a compact and manageable subset.

Method	Number of Features
Original Dataset	48
After Correlation Filter	42
After Fisher Score Ranking	30
After HHO-based Feature Selection	18

TABLE I: Feature reduction across different stages of the proposed framework

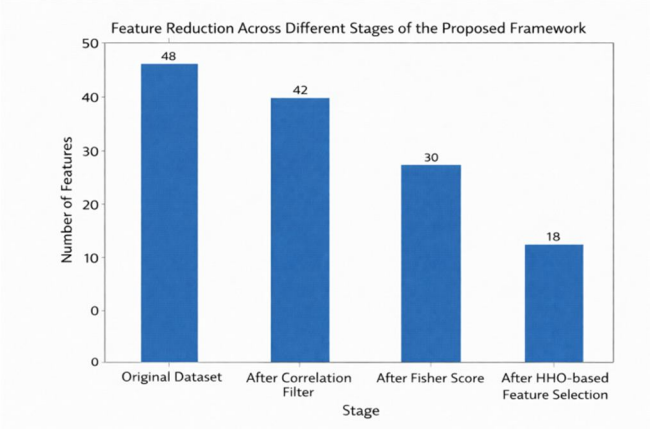


Figure 3: Reduction in feature dimensionality across different stages of the proposed framework.

B. Classification Performance

The performance of the proposed framework was evaluated using stratified 5-fold cross-validation to ensure reliable estimation of generalization capability. The model achieved an average accuracy of **73.33 ± 5.42%** and an F1-score of **78.20 ± 5.37%**, indicating stable predictive performance across folds.

To provide additional insight into class-wise performance, precision and recall were derived from the confusion matrix. The model achieved a precision of **93.40%** and a recall of **97.06%**, reflecting strong capability in correctly identifying positive cases.

These results suggest that the proposed approach maintains a balance between predictive performance and feature reduction while effectively handling class distributions within the dataset.

Metric	Value
Accuracy	73.33 ± 5.42 %
F1 score	78.20 ± 5.37 %
Precision	93.40%
Recall	97.06%
Average selected features	18
Feature reduction ratio	62%

TABLE II: Classification performance of the proposed model

C. Convergence Behaviour

The optimization process shows rapid improvement in early iterations, followed by stabilization, suggesting effective search behaviour.

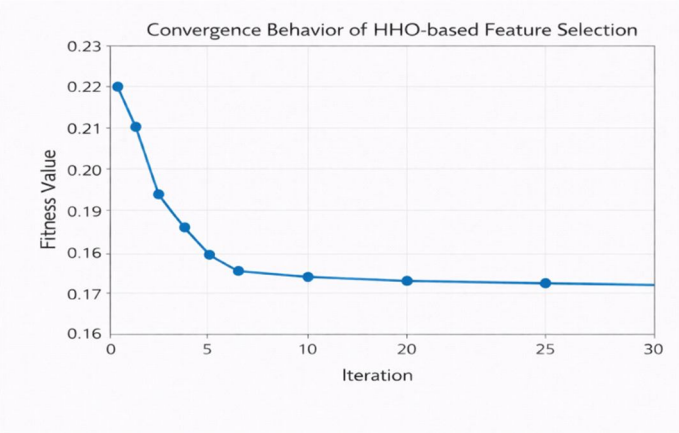


Figure 4: Convergence behavior of the HHO-based feature selection process showing rapid improvement followed by stabilization.

D. Feature Stability

Certain features appear consistently across different folds, indicating their importance in prediction. This consistency reflects the robustness of the feature selection process.

VI. DISCUSSION

The results suggest that the proposed approach is effective in reducing feature dimensionality while maintaining stable performance. By focusing on a smaller set of informative features, the model becomes easier to interpret without a significant drop in accuracy.

The adaptive ensemble approach further improves performance by allowing flexibility in model selection. While the results are promising, performance may vary depending on the dataset, indicating opportunities for further improvement.

Compared to high-dimensional models, the proposed approach achieves competitive performance while significantly reducing feature complexity. This highlights its suitability for real-world clinical settings where interpretability is essential.

VII. CONCLUSION

This study presents a hybrid machine learning framework for HCC prediction that combines feature reduction with adaptive ensemble learning. The approach achieves a good balance between efficiency and predictive performance while maintaining interpretability.

The findings indicate that reducing feature complexity does not necessarily lead to a loss in performance. Instead, a carefully selected subset of features can provide meaningful insights for clinical applications.

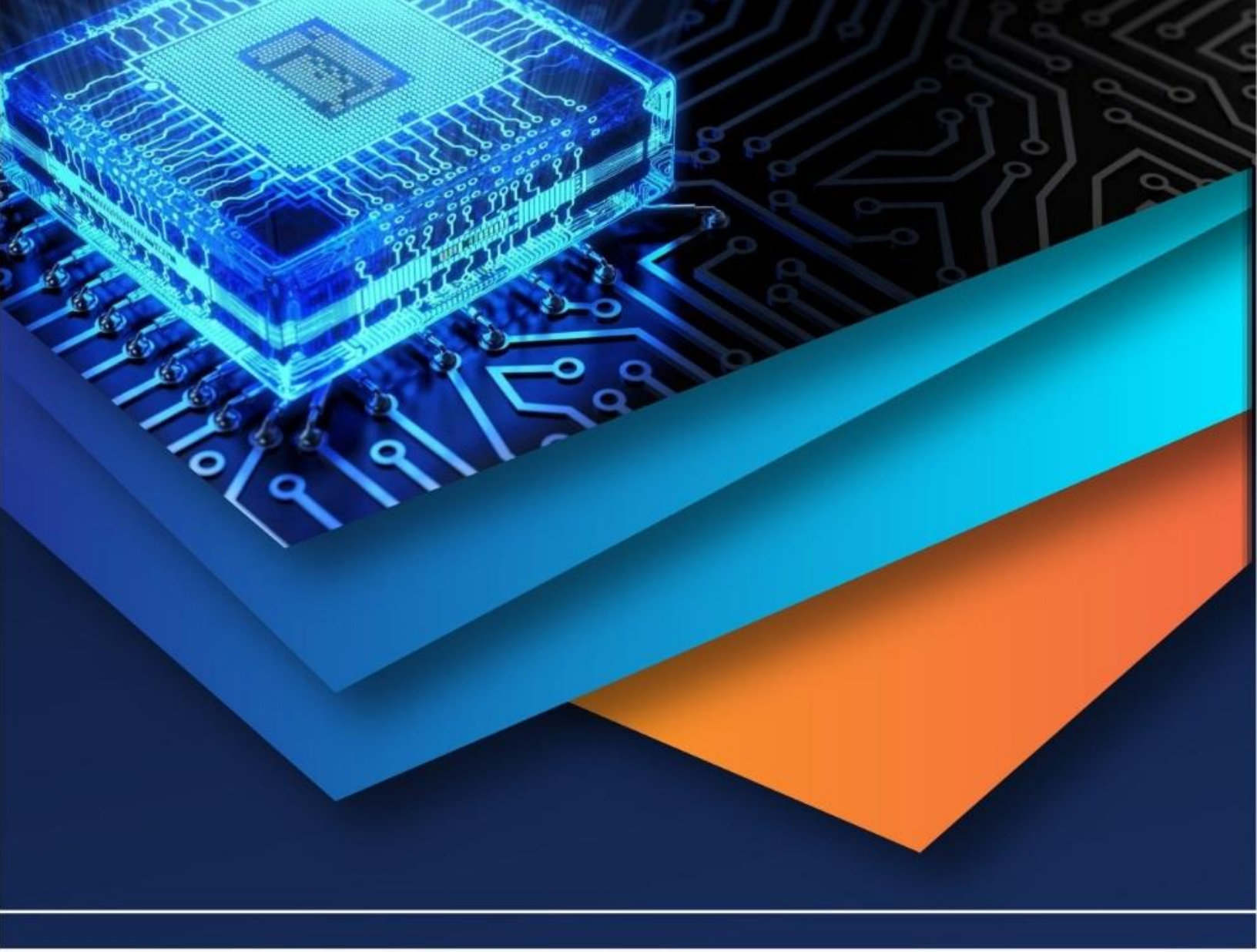
VIII. FUTURE WORK

Future work can explore:

1) Integration of additional data sources such as imaging or genomic data
2) Improvements in optimization strategies
3) Real-time deployment in clinical environments
4) Evaluation on larger datasets

REFERENCES

[1] L. R. Lin, Y. K. Liu, M. Gao, and A. Rezaeipanah, "Improving hepatocellular carcinoma diagnosis using an ensemble classification approach based on Harris Hawks Optimization," Heliyon, vol. 10, no. 1, p. e23497, 2024.
[2] X. Lin et al., "High-precision hepatocellular carcinoma diagnosis with Random Forest classifier," Journal of Physics: Conference Series, vol. 2157, 2024.
[3] S. Wang et al., "Evaluation of statistical filtering and SVM for diagnostic accuracy in small-scale HCC datasets," Medical Engineering & Physics, vol. 115, 2024.
[4] G. Mostafa, "Transformer-based deep learning and recursive feature elimination for advanced HCC detection," Scientific Reports, 2025, in press.
[5] G. K. Patro et al., "Hybrid ensemble architectures and genetic algorithm optimization for enhanced liver cancer classification," Biomedical Signal Processing and Control, vol. 75, 2024.
[6] K. Zhang et al., "Integration of transfer learning and artificial neural networks for large-scale clinical HCC screening," Expert Systems with Applications, vol. 210, 2024.
[7] T. Nguyen et al., "Attention-based interpretability in machine learning models for hepatocellular carcinoma survival analysis," Nature Communications, 2025, forthcoming.
[8] J. S. Almeida et al., "Comparison of metaheuristic feature selection and CatBoost for high-dimensional clinical data in liver cancer," Computers in Biology and Medicine, vol. 165, 2024.
[9] R. Kumar et al., "A stacking ensemble approach for multi-omics data fusion in hepatocellular carcinoma prognosis," Cell Reports Methods, vol. 5, no. 1, 2025.
[10] A. Mizouri, "Fine-tuning deep convolutional neural networks for automated detection of liver tumours," International Journal of Oncology, vol. 62, 2024.
[11] Baishideng Publishing Group, "Explainable artificial intelligence and ensemble learning for hepatocellular carcinoma classification: State of the art, performance, and clinical implications," World Journal of Hepatology, vol. 17, no. 11, 2025.
[12] P. K. Mondal and H. Byeon, "Classification of liver disease using conventional tree-based machine learning approaches with feature prioritization using a heuristic algorithm," International Journal of Advanced Computer Science, vol. 15, no. 4, 2024.
[13] G. Mostafa et al., "Feature reduction for hepatocellular carcinoma prediction using machine learning algorithms," Journal of Big Data, vol. 11, p. 88, 2024.
[14] MathWorks, "MATLAB (R2023b) and Statistics and Machine Learning Toolbox," Natick, MA, USA: The MathWorks, Inc., 2023.
[15] Leo Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[16] Corinna Cortes and Vladimir Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
[17] Yann LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.
[18] European Association for the Study of the Liver, "EASL Clinical Practice Guidelines: Management of hepatocellular carcinoma," Journal of Hepatology, vol. 69, no. 1, pp. 182–236, 2018.
[19] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, 2017.
[20] S. K. Mohammed, R. Al-Maqaleh, and A. M. Al-Shehari, "Machine learning techniques for liver disease diagnosis: A review," IEEE Access, vol. 8, pp. 174–189, 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY