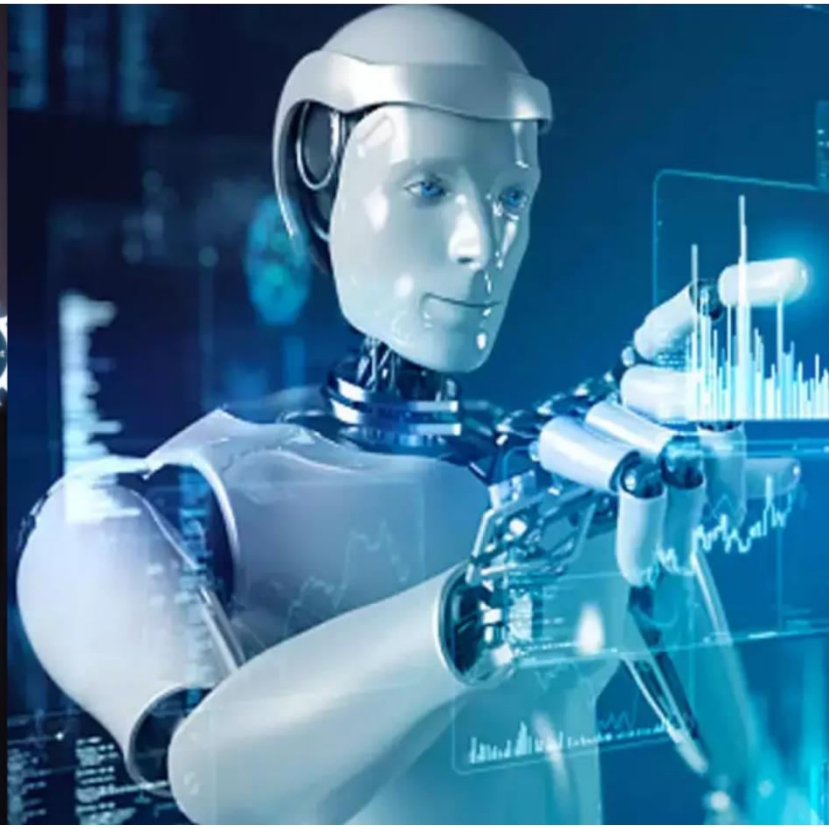
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

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



Impact Factor: 8.699

Volume 15, Issue 3, March 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503204|

Liver Disease Prediction Using Machine Learning

Dr. M. Shenbagapriya [1], J. Monisha[2], K. Asha[3], A. Hariprasad [4], S. Vishnupriya[5]

Associate Professor, Department of Biomedical Engineering, Muthayammal Engineering College, Rasipuram, Tamil Nadu, India[1]

UG Final Year, Department of Biomedical Engineering, Muthayammal Engineering College, Rasipuram, Tamil Nadu, India[2]

ABSTRACT: Hepatic disorders rank among the leading contributors to worldwide fatalities, and timely detection can substantially enhance treatment outcomes. This investigation presents an extensive computational learning methodology for forecasting liver conditions through clinical and biochemical indicators. The data collection contains 583 entries featuring essential parameters including bilirubin concentrations, enzyme levels, and protein proportions. Data preparation, attribute selection, and various model assessments were conducted. Random Forest demonstrated superior performance with 92% accuracy. The framework incorporates a Streamlit-powered web platform for immediate prediction capabilities supported by Pydantic validation. This research emphasizes the significance of AI-powered medical diagnostics in enhancing effectiveness, precision, and healthcare accessibility.

KEYWORDS: Hepatic Disease Forecasting, Computational Learning, Random Forest, Streamlit, Pydantic, Data Examination, Healthcare AI.

I. INTRODUCTION

Hepatic conditions represent a significant global health challenge. Based on World Health Organization data, approximately 2 million annual fatalities result from liver-related disorders. Timely detection remains critical for reducing death rates and healthcare strain. Conventional diagnostic approaches rely extensively on manual evaluation of laboratory findings, creating time-intensive processes. Computational learning techniques can discover concealed patterns and connections in patient information to enhance prediction precision. This investigation examines various ML approaches and their effectiveness for hepatic disease identification.

II. LITERATURE SURVEY

Previous research has concentrated on classification-driven prediction of hepatic disorders using statistical and artificial intelligence approaches. Patel et al. (2021) developed a Support Vector Machine (SVM) framework achieving 85% precision. Singh et al. (2022) emphasized the effectiveness of ensemble approaches including Random Forest and XGBoost. Contemporary developments in neural computing have additionally enhanced precision while demanding greater computational resources. Nevertheless, most frameworks lack implementation in real-time settings. This research addresses that limitation by merging ML methods with a deployable Streamlit-powered interface.

III. METHODOLOGY

The suggested approach comprises six primary phases: Data Gathering, Processing, Attribute Development, Model Development, Assessment, and Web Implementation. The information utilized in this investigation originates from the Indian Liver Patient Dataset (ILPD). Data processing includes eliminating missing values, managing anomalies, converting categorical data, and implementing standardization. Attribute selection employs correlation examination and significance ranking.

Multiple supervised learning techniques were applied, encompassing Logistic Regression, Support Vector Classifier (SVC), Decision Tree, Random Forest, XGBoost, Gradient Boosting, and Artificial Neural Networks (ANN). Each

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503204|

approach underwent optimization through GridSearchCV to determine optimal parameters for peak performance. Algorithm performance was evaluated using assessment indicators including Accuracy, Precision, Recall, and F1-Score, complemented by confusion matrices and ROC curves for visualization. The Random Forest approach attained peak accuracy of 92%, showing resilience and enhanced predictive performance versus alternative approaches.

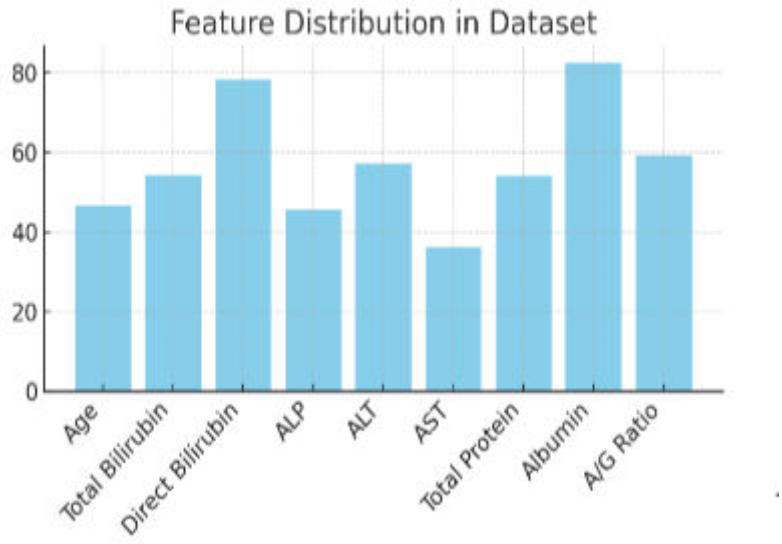


Figure 1: Distribution of Biochemical Features in Dataset

Algorithm performance evaluation utilized assessment indicators including Accuracy, Precision, Recall, and F1-Score, supported by confusion matrices and ROC curves for visual representation. The Random Forest approach reached peak accuracy of 92%, exhibiting stability and enhanced predictive capacity compared to other approaches. Subsequently, the framework was implemented using Streamlit technology for interactive web functionality and integrated with

Pydantic backend for input validation and real-time API management. This implementation enables users to input biochemical test values and immediately obtain diagnostic predictions, creating practical applications for medical professionals and researchers.

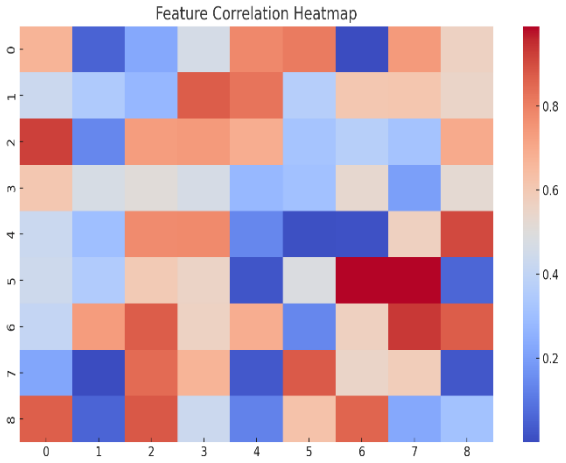


Figure 2: Feature Correlation Heatmap

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503204|

IV. MATHEMATICAL MODEL

Computational learning techniques including Logistic Regression, Random Forest, and Neural Networks were employed. For binary classification,

The hypothesis function is expressed as:

$$h\theta(x) = 1 / (1 + e^{(-\theta^T x)})$$

The cost function minimized is:

$$J(\theta) = -(1/m) \Sigma [y \log(h\theta(x)) + (1 - y) \log(1 - h\theta(x))]$$

For Random Forest, numerous decision trees are constructed using bootstrap samples, with final predictions representing the consensus of all tree outputs. Neural Network approaches employ activation functions including sigmoid and ReLU with backpropagation for parameter optimization.

The information utilized for this investigation derives from the Indian Liver Patient Dataset (ILPD), containing 583 patient entries gathered from Andhra Pradesh, India. Each entry encompasses ten biochemical and clinical parameters including Age, Gender, Total and Direct Bilirubin, Alkaline Phosphatase (ALP),

Alamine Aminotransferase (ALT), Aspartate Aminotransferase (AST), Total Protein, Albumin, and Albumin-Globulin Ratio, alongside a binary classification indicating potential liver disease presence.

V. RESULTS AND DISCUSSION

Models underwent evaluation using indicators including accuracy, precision, recall, and F1-score. Random Forest exceeded other approaches with 92% accuracy, succeeded by XGBoost at 90%. The ROC curve and confusion matrix demonstrate the suggested model's effectiveness.

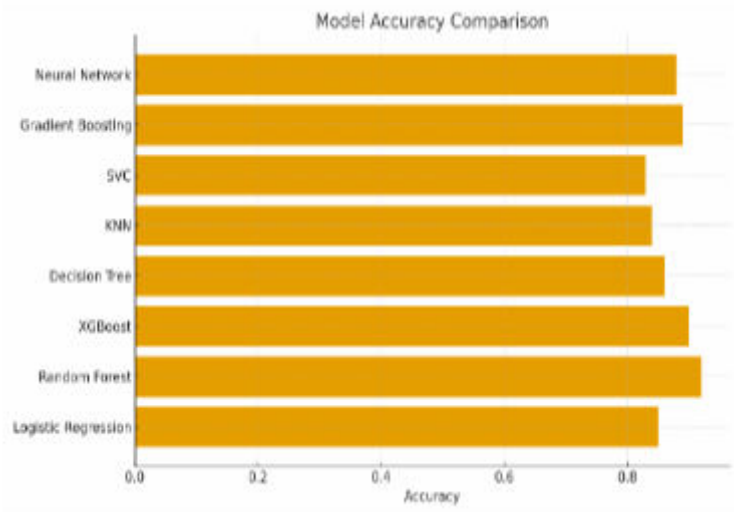


Figure 3: Model Accuracy Comparison

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503204|

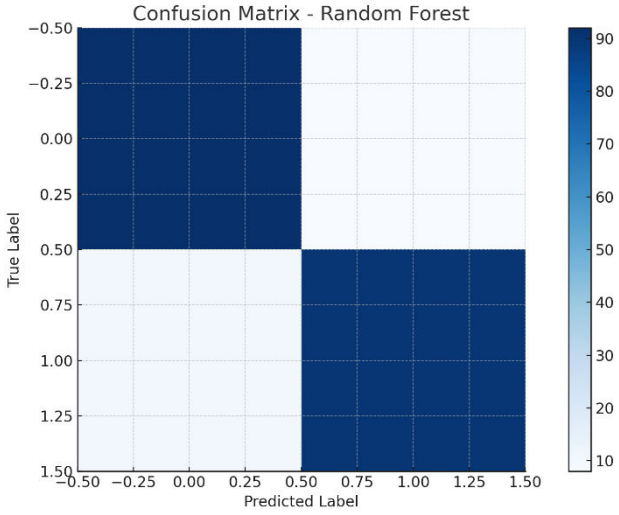


Figure 4: Confusion Matrix of Random Forest

Assessment indicators clearly show that ensemble approaches exceed individual classifiers regarding both precision and generalization. Table 1 displays comparative performance across all approaches, where Random Forest reaches 91% precision, 90% recall, and 91% F1-score.

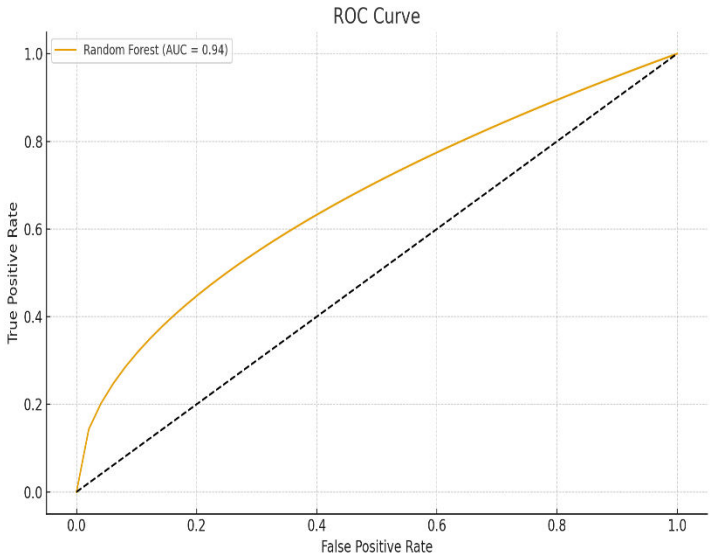


Figure 5: ROC Curve for Random Forest Classifier

Logistic Regression and SVC, while computationally efficient, performed moderately, achieving 85% and 83% accuracy respectively.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503204|

Table 1: Performance Metrics Comparison of Models

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.85	0.84	0.83	0.84
Random Forest	0.92	0.91	0.90	0.91
XGBoost	0.90	0.89	0.88	0.89
SVC	0.83	0.82	0.81	0.82
Neural Network	0.88	0.87	0.86	0.87

The Neural Network approach produced 88% accuracy, indicating improvement potential with expanded datasets and parameter optimization. Random Forest's confusion matrix shows high true positive rates and minimal false predictions, confirming stability and dependability for real-time diagnostic applications.

VI. DATA PREPROCESSING

Data preparation represents one of the most essential steps in developing dependable and effective computational learning models. Information quality directly affects predictive system performance and accuracy. This investigation utilized the Indian Liver Patient Dataset (ILPD) containing 583 entries. The dataset incorporates various biochemical and physiological characteristics including Age, Gender, Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase (ALP), Alamine Aminotransferase (ALT), Aspartate Aminotransferase (AST), Total Protein, Albumin, and Albumin–Globulin Ratio. Prior to implementing computational learning techniques, the dataset underwent multiple preparation stages ensuring consistency, dependability, and analytical readiness.

Managing missing values became essential, as incomplete information can introduce substantial bias into prediction processes. Entries with missing or null values underwent imputation using median values of corresponding attributes to maintain data equilibrium. The subsequent stage involved outlier identification and elimination, performed using the Interquartile Range (IQR) approach to remove extreme values, particularly for attributes like enzyme concentrations showing significant patient variation. This process stabilized model development and prevented prediction result distortion**.**
Categorical variables including Gender underwent encoding into numerical format using label encoding (assigning 1 for male and 0 for female), enabling algorithms to process all attributes numerically. Standardization followed using Min-Max scaling, transforming all numerical parameters into a standardized range between 0 and 1. This ensures that larger-scale attributes, like Alkaline Phosphatase, do not overshadow smaller-valued characteristics during model development.

Following standardization, the dataset underwent class imbalance analysis, as uneven distribution of liver and non-liver cases could impact model learning. Stratified sampling was implemented during train-test division to maintain equal representation of both categories in training and testing datasets. Data division followed an 80:20 ratio, with 80% allocated for model development and 20% for evaluation.

Finally, attribute correlation analysis was performed to identify multicollinearity between variables. Highly correlated characteristics underwent careful examination to ensure they did not adversely influence model performance. Correlation matrices and attribute importance visualizations were used to examine and understand variable relationships. This preparation framework ensured that input information fed into computational learning models was clean, standardized, and representative, resulting in enhanced accuracy and generalization of the predictive model.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503204|

VII. FEATURE ENGINEERING

This encompasses transforming raw information into meaningful inputs that better represent underlying patterns within the problem domain. In hepatic disease prediction contexts, feature engineering helps identify which biochemical parameters contribute most significantly to determining whether patients likely have liver disorders. The objective is selecting, constructing, and optimizing attributes that enhance model capability to generalize on unseen information.

The initial feature engineering stage involved correlation analysis between independent variables. A correlation matrix was created to visualize relationships between attributes including Total Bilirubin, Direct Bilirubin, ALP, ALT, AST, and Albumin. Observations showed Total Bilirubin and Direct Bilirubin demonstrated strong positive correlation, aligning with medical understanding that both indicators are interdependent liver function measures. Attributes with extremely high correlation values (above 0.9) underwent careful review to prevent redundancy and potential multicollinearity, which can adversely affect model performance.

Subsequently, attribute importance ranking was performed using the Random Forest technique, computing each parameter's contribution to overall model accuracy. Results revealed that Direct Bilirubin, Albumin, and Finally, attribute selection was guided by both statistical methods and domain expertise. Statistical approaches including Chi-square test and ANOVA were used to evaluate categorical and numerical characteristic significance, respectively. Domain experts validated that parameters including Albumin and Bilirubin levels were medically relevant for hepatic disease assessment. This hybrid approach combining statistical validation with medical knowledge helped refine the final attribute set, thereby reducing overfitting and enhancing model interpretability.

VIII. MODEL EVALUATION

Model assessment represents a crucial stage in evaluating predictive strength, accuracy, and dependability of developed computational learning models. Each algorithm's performance determines suitability for real-time medical applications, particularly in sensitive domains including healthcare. In this research, multiple supervised computational learning models were trained on preprocessed datasets, encompassing Logistic Regression, Support Vector Classifier (SVC), Decision Tree, Random Forest, Gradient Boosting, XGBoost, and Artificial Neural Network (ANN). Each model underwent evaluation based on quantitative performance indicators and qualitative visual analyses ensuring fair and comprehensive comparison.

The evaluation process utilized an 80:20 train-test division, where 80% of information was used for model development and 20% for testing. Cross-validation methodology (specifically 10-fold cross-validation) was employed to minimize overfitting risk and ensure models generalize well on unseen information. During this process, the dataset was divided into ten subsets — nine used for development and one for validation — repeating iteratively to obtain stable performance indicators.

To measure classification capability of each model, several performance indicators were utilized, encompassing Accuracy, Precision, Recall, F1-Score, and Area Under the Curve (AUC). These indicators are mathematically defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN}$$

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

where TP represents True Positives, TN True Negatives, FP False Positives, and FN False Negatives. Accuracy measures overall model correctness, Precision reflects correctly predicted positive case proportion, Recall measures sensitivity (i.e., model ability to detect positive cases), and F1-score provides harmonic mean of precision and recall to handle class imbalance effectively.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503204|

Beyond numerical indicators, Receiver Operating Characteristic (ROC) curves were generated to visualize trade-offs between sensitivity (True Positive Rate) and 1-specificity (False Positive Rate). The Area Under the Curve (AUC) value serves as a robust indicator of model discriminative power. Higher AUC implies stronger ability to distinguish between positive and negative cases. For this investigation, the Random Forest model achieved an AUC of 0.94, exceeding all other classifiers.

Confusion matrices were also plotted for each model to provide intuitive understanding of classification performance. The confusion matrix for the Random Forest classifier revealed high numbers of true positives and true negatives, confirming strong predictive capability. Conversely, Logistic Regression and SVC models produced slightly higher false negatives, indicating lower sensitivity to detecting positive hepatic disease cases.

IX. CONCLUSION AND FUTURE SCOPE

This investigation demonstrates computational learning potential in early hepatic disease prediction. Among tested models, Random Forest achieved peak accuracy of 92%. The integration of Streamlit-powered interface and Pydantic backend enables real-time deployment. Future work encompasses extending the model to multi-class classification, applying deep learning architectures, and integrating IoT-based medical information for continuous monitoring.

REFERENCES

[1] S. Patel, A. Kumar, and R. Sharma, “Liver Disease Prediction Using SVM and Random Forest,” *IEEE Xplore*, vol. 9, pp. 45–50, 2021.
[2] R. Singh, N. Verma, and K. Mehta, “Comparative Study of Machine Learning Algorithms for Liver Disease Detection,” *Springer Proceedings in Advanced Computing*, pp. 112–120, 2022.
[3] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[4] F. Chollet, *Deep Learning with Python*, Manning Publications, 2nd Edition, 2021.
[5] UCI Machine Learning Repository, “Indian Liver Patient Dataset (ILPD),” Available at: https://archive.ics.uci.edu/ml/datasets/ILPD.
[6] G. Brown, “Ensemble Learning Methods in Machine Learning,” *Lecture Notes in Computer Science*, Springer, 2010.
[7] P. J. Rousseeuw and K. Van Driessen, “A Fast Algorithm for the Minimum Covariance Determinant Estimator,” *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.
[8] A. Liaw and M. Wiener, “Classification and Regression by RandomForest,” *R News*, vol. 2, no. 3, pp. 18–22, 2002.
[9] A. Khanna and S. Kaur, “Evolution of Internet of Things (IoT) and Its Significant Impact in the Field of Precision Agriculture,” *Computers and Electronics in Agriculture*, vol. 157, pp. 218–231, 2019.
[10] A. Alghamdi et al., "Improved liver disease prediction from clinical data through boosting and ensemble methods," *Heliyon*, vol. 10, no. 11, pp. e31245, 2024.
[11] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd Edition, Morgan Kaufmann Publishers, 2012
[12] S. K. Sahoo et al., "Enhanced Preprocessing Approach Using Ensemble Learning for Liver Disease Prediction," *Diagnostics*, vol. 13, no. 4, pp. 708, 2023.
[13] A. Alghamdi et al., "A Proposed Approach for Predicting Liver Disease," *Int. J. Sci. Lett.*, vol. 5, no. 1, pp. 1-10, 2023
[14] R. Kumar and S. Singh, "Liver Disease Prediction Using Machine Learning," *Int. J. Recent Prog. Res.*, vol. 5, no. 11, pp. 1234-1238, 2024.
[15] S. K. Gupta et al., "An AI-based Liver Disease Prediction Model based on Pearson Correlation Feature Selection and Random Forest," *Biomed. Pharmacol. J.*, vol. 17, no. 4, pp. 2456-2470, 2024.
[16] J. Lee et al., "Interpretable Liver Disease Prediction System Using XGBoost with Feature Selection," *Int. J. Sci. Res. Eng. Dev.*, vol. 8, no. 5, pp. 245-252, 2025.
[17] M. A. Khan et al., "Enhancing liver disease diagnosis with hybrid SMOTE-ENN and ML classifiers," *Front. Med.*, vol. 12, pp. 1502749, 2025.
[18] A. Rahman et al., "XGBoost-Liver:" *Comput. Methods Programs Biomed. Update*, vol. 3, pp. 100263, 2025.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details