IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155137

AI-Based Early Detection and Risk Prediction of Jaundice Using Clinical and Liver Function Test Data

Vaseekaran A[1], Hariharan S[2], and Mrs. Malathi G[3]

Final Year Student, Department of Artificial Intelligence and Data Science, Dhanalakshmi Srinivasan College of Engineering and Technology, Affiliated to Anna University, Chennai 600 025, Tamil Nadu, India[1,2]

Assistant Professor, Department of Artificial Intelligence and Data Science, Dhanalakshmi Srinivasan College of Engineering and Technology, Affiliated to Anna University, Chennai 600 025, Tamil Nadu, India[3]

Abstract: Jaundice, the yellow discolouration of skin and sclera caused by elevated serum bilirubin, is a clinically visible manifestation of hepatic, haematological or biliary dysfunction whose early detection materially affects patient outcomes. In the Indian subcontinent, where viral hepatitis A and E are endemic and non-alcoholic fatty liver disease is rising in prevalence, the burden of liver disease is substantial while specialist hepatology expertise is concentrated in tier-one urban centres. This paper presents HepatIQ, an artificial-intelligence-driven decision support system for the early detection and risk stratification of jaundice. The system combines a Random Forest classifier trained on the Indian Liver Patient Dataset of 583 records with a rule-based pattern classifier and a biochemical flagging engine to produce explainable risk assessments. The system achieves a five-fold cross-validation accuracy of 70.68% (±3.14%) and a test-set accuracy of 72.65%, with a clinically critical recall of 100% on the high-risk class at 97% precision. The hybrid combination of probabilistic machine learning output, pattern classification, biochemical flags and India-specific dietary recommendations addresses the explainability gap that limits adoption of black-box predictors in clinical settings. The complete system is delivered as a Flask-based web application with SQLite persistence, runs on commodity hardware, and uses only open-source libraries.

Index Terms: Jaundice, Liver Function Test, Random Forest, Machine Learning, Indian Liver Patient Dataset, Risk Stratification, Clinical Decision Support, Explainable AI, Hepatology, Bilirubin.

I. INTRODUCTION

THE human liver performs more than five hundred distinct metabolic, secretory, detoxification and synthetic functions, including the conjugation and excretion of bilirubin formed during normal red blood cell turnover. When this orderly pathway is disrupted at any point, bilirubin accumulates in body tissues and produces the visible yellow discolouration recognised clinically as jaundice or icterus [1]. Jaundice is not itself a disease but a sign whose underlying causes range from benign Gilbert syndrome to potentially fatal fulminant hepatic failure, biliary obstruction and hepatocellular carcinoma.

According to the World Health Organization, viral hepatitis was associated with approximately 1.3 million deaths globally in 2022, exceeding mortality from human immunodeficiency virus and tuberculosis combined [2]. The Indian subcontinent bears a disproportionate share of this burden owing to the endemic presence of hepatitis A and E, large pockets of inadequate sanitation, rising prevalence of obesity-related fatty liver disease, high rates of alcohol-related cirrhosis in certain demographic groups, and the use of unregulated traditional remedies with hepatotoxic ingredients. Studies published in the Journal of Clinical and Experimental Hepatology have estimated the prevalence of non-alcoholic fatty liver disease in urban Indian adults at 25-40% [3], with silent progression to non-alcoholic steatohepatitis, fibrosis, cirrhosis and hepatocellular carcinoma in a worrying proportion of cases.

The standard diagnostic workup for suspected jaundice begins with the liver function test (LFT) panel, which measures serum concentrations of total bilirubin, direct bilirubin, alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), gamma-glutamyl transferase (GGT), total proteins, albumin and the albumin-to-globulin ratio. The interpretation of these multi-parameter results requires substantial clinical experience, as the patterns of derangement reveal the most likely site and mechanism of injury. Disproportionate elevations of ALT and AST suggest hepatocellular injury; elevations dominated by ALP and GGT suggest cholestatic patterns; isolated bilirubin elevations

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155137

with normal enzymes suggest pre-hepatic causes such as haemolysis [4]. In primary and secondary healthcare settings, where specialist hepatologists are unavailable and physicians may see seventy or eighty patients per morning clinic, the consistent and timely interpretation of these complex panels remains a significant challenge.

Artificial intelligence (AI) and machine learning (ML) have emerged over the past decade as transformative tools for clinical decision support. Modern supervised learning algorithms can identify patterns in high-dimensional clinical data that escape human cognition, deploy at scale and at low cost, and provide consistent, reproducible outputs free from the inter-observer variability that troubles human interpretation [5]. Applications to hepatology have been explored extensively in recent literature, with classifiers trained on LFT panels, ultrasound images, magnetic resonance elastography data and electronic health records to predict liver fibrosis, hepatocellular carcinoma, transplant outcomes and acute liver injury severity [6], [7].

Despite this progress, three persistent gaps limit the clinical impact of published work. First, most published ML systems for liver disease prediction conclude with cross-validated metrics on a static dataset, leaving the work of system integration, user interface design and database persistence unaddressed. Second, the explainability layer recognised as essential for clinical adoption is often discussed in the abstract but rarely implemented in a deployable manner [8]. Third, India-specific contextualisation — covering dietary recommendations, monsoon-season hepatitis epidemiology, and the baseline biochemistry of vegetarian and non-vegetarian populations — is seldom delivered in published academic work.

A. Contributions of This Work

This paper presents HepatIQ, a complete end-to-end clinical decision support system for the early detection and risk stratification of jaundice. The principal contributions are as follows:

1) A hybrid prediction architecture that combines a class-balanced Random Forest classifier with a rule-based pattern classifier and a biochemical flagging engine, addressing the explainability gap that limits adoption of pure black-box predictors in clinical settings.
2) A label engineering scheme that converts the original binary disease label of the Indian Liver Patient Dataset into a clinically meaningful three-class severity scale based on bilirubin thresholds, supporting consistent triage decisions.
3) A five-category user-facing risk scale (low, mild, moderate, high, critical) mapped from the underlying three-class probabilities through a weighted risk-score formula, enabling more granular clinical communication.
4) A complete reference implementation as a Flask-based web application with SQLite persistence, demonstrating that AI-driven decision support can be deployed on commodity hardware using only open-source libraries.
5) India-specific dietary, lifestyle and follow-up recommendations adapted to local food habits and the regional epidemiology of viral hepatitis.

On evaluation, the system achieves 70.68% (±3.14%) five-fold cross-validation accuracy, 72.65% test-set accuracy and, most importantly from a clinical safety standpoint, 100% recall on the high-risk class at 97% precision — ensuring that no severely compromised patient is overlooked.

B. Public Health Significance

The public health significance of jaundice detection in the Indian context cannot be overstated. India has one of the highest burdens of viral hepatitis in the world, with an estimated 40 million chronic hepatitis B carriers and 6-12 million chronic hepatitis C cases. Acute viral hepatitis A and E cause seasonal outbreaks particularly during the monsoon months of June through September, when waterborne transmission peaks in regions with poor sanitation infrastructure. The Indian Council of Medical Research has repeatedly identified liver disease as among the top ten causes of years of life lost in working-age adults.

The disparity between urban and rural healthcare access exacerbates this burden. Tier-one metropolitan cities such as Delhi, Mumbai, Chennai and Bangalore have a reasonable density of hepatologists and gastroenterologists, but the ratio of specialists to population in tier-two and tier-three cities, towns and villages is often worse than one per hundred thousand inhabitants. The result is that the overwhelming majority of patients with suspected liver disease in India are evaluated initially — and often exclusively — by general physicians, primary health centre doctors or trained medical officers whose training in interpreting complex LFT patterns is variable. An accessible, automated decision support tool with India-specific calibration thus has the potential to make a measurable contribution to population-level health outcomes by reducing diagnostic delays, improving triage consistency and freeing specialist time for the most complex cases.

C. Why Machine Learning?

Machine learning offers several specific advantages over rule-based expert systems for the interpretation of LFT panels. First, learned models can capture non-linear interactions between biomarkers that are difficult to express as explicit rules — for example, the changing significance of moderately elevated ALT depending on the simultaneous values of bilirubin, ALP and albumin. Second, learned models adapt to the actual population distribution of the training data rather than

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155137

relying on textbook reference ranges that may not reflect the baseline biochemistry of the local population. Third, learned models can be retrained as new data accumulate, allowing the system to improve over time without requiring manual revision of explicit rules. Fourth, learned models provide native probability estimates that quantify uncertainty in a way that crisp rules cannot, supporting nuanced clinical communication around marginal cases.

These advantages are tempered, however, by the explainability problem: the internal reasoning of a learned model is opaque to its users, which severely limits clinician acceptance. The present work addresses this tension by combining the learned model with rule-based components that surface the underlying reasoning, an architectural choice that retains the strengths of both paradigms while mitigating their respective weaknesses.

D. Paper Organisation

The remainder of this paper is organised as follows. Section II reviews related work in machine learning applications to liver disease prediction. Section III describes the dataset and the preprocessing pipeline. Section IV presents the proposed system architecture. Section V details the model design including the random forest configuration and the hybrid prediction engine. Section VI outlines the implementation. Section VII presents experimental results and a comparative analysis. Section VIII presents ablation and sensitivity studies. Section IX discusses the clinical implications and limitations. Section X concludes the paper and identifies directions for future work.

II. RELATED WORK

A. Machine Learning for Liver Disease Prediction

The application of machine learning to liver disease prediction has been a subject of active investigation. Khanna et al. [9] presented a comprehensive comparative study of seven supervised algorithms — logistic regression, k-nearest neighbours, support vector machines with linear and RBF kernels, decision trees, random forests, gradient boosting and XGBoost — for the prediction of liver disease using the Indian Liver Patient Dataset (ILPD). The authors reported that ensemble methods consistently outperformed single-model classifiers, with random forest and XGBoost achieving accuracies of approximately 72% and 74% respectively after careful hyperparameter tuning. A particular strength of their work was the treatment of class imbalance through both SMOTE oversampling and class-weighted loss functions, with the authors concluding that class-balanced random forest offers the best practical compromise between performance, interpretability and computational cost.

Ramachandran et al. [10] applied a three-layer feed-forward neural network with batch normalisation and dropout regularisation to a multi-centre Indian dataset of 2,000 patients, achieving 78% accuracy in cirrhosis prediction. However, the authors observed that the deep model required substantially more training data and computational resources than ensemble tree-based methods, with only modest gains in performance. They explicitly recommended classical ensemble methods for datasets in the ILPD size range, a recommendation that directly informed the algorithmic choice in the present work.

Reddy et al. [11] conducted a systematic comparison of four tree-based ensemble algorithms — random forest, gradient boosting, AdaBoost and XGBoost — on twenty medical classification benchmarks. They found that random forest consistently produced the most stable performance with the lowest variance across cross-validation folds, while XGBoost achieved higher peak accuracy but with greater sensitivity to hyperparameter choices. The authors recommended random forest for medical applications where stability, interpretability and robustness to varied data distributions outweigh peak benchmark accuracy.

B. Explainable AI in Clinical Decision Support

Sharma and Patel [12] addressed the problem that black-box machine learning models, despite high predictive performance, are seldom adopted in clinical practice owing to physician concerns about transparency, accountability and liability. The authors developed a hybrid framework combining a gradient-boosting classifier with SHAP (Shapley Additive Explanations) values and rule-based reasoning derived from clinical practice guidelines. Their evaluation on a multi-centre Indian dataset showed that physicians presented with both predictions and accompanying explanations were 93% more likely to act on the model's recommendations than those who saw only raw scores.

Fernandez and Kumar [13] presented a theoretical and empirical analysis of hybrid systems combining ML predictions with rule-based reasoning. They derived a taxonomy of integration patterns including pre-filtering, post-filtering, parallel-with-arbitration and weighted-fusion approaches, concluding that for diagnostic decision support the parallel-with-arbitration pattern offers the best balance between performance and explainability. This is precisely the integration pattern adopted in the present work.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155137

C. Symptom Integration and Clinical Context

Krishnamoorthy et al. [14] demonstrated that the addition of clinical symptoms to a biochemical-only feature set improved acute viral hepatitis prediction by an average of nine percentage points across multiple classifiers. The symptoms found most predictive included scleral icterus, dark urine, pale stools, fatigue and right upper quadrant abdominal pain. This finding supported the design choice of including nine clinical symptom features alongside eight LFT parameters in the present work.

Joshi et al. [15] presented a clinical severity scoring scheme for jaundice based on bilirubin thresholds, aminotransferase levels and clinical signs. Their three-tier classification of mild, moderate and severe disease, validated on a cohort of 1,200 Indian patients, demonstrated good correlation with subsequent clinical outcomes. The Joshi scoring scheme directly informed the design of the three-class label engineering used in the present work, with the bilirubin threshold of 2.0 mg/dL chosen as the discriminating cutoff.

D. Class Imbalance and Sampling Strategies

Singh and Banerjee [16] conducted a thorough empirical study of the impact of class imbalance on random forest performance in twelve benchmark medical datasets including the ILPD. They demonstrated that the class_weight=balanced parameter in scikit-learn improved minority-class recall by an average of fourteen percentage points without significantly degrading precision. They also showed that combining class weighting with stratified k-fold cross-validation provided the most reliable estimates of real-world performance. Both techniques have been adopted in the present work.

E. Web-Based Clinical Decision Support

Iyengar and Reddy [17] surveyed seventeen web-based clinical decision support systems for hepatology described in the literature between 2018 and 2024. They reported that fewer than 30% of these systems remained operational at the time of survey, primarily due to poor maintenance, dependency on discontinued commercial cloud services and the absence of incremental retraining mechanisms. The authors recommended a set of architectural principles for sustainable deployment including the use of open-source technologies, local data persistence, modular design and documented retraining procedures. These principles have been adhered to in the design of HepatIQ.

Verma et al. [18] examined the use of Flask for healthcare web applications and identified the framework as particularly well-suited for projects in the one-to-twenty-thousand-user range due to its minimalist design, low memory footprint, and rich Python data-science ecosystem integration. They noted that for very high traffic deployments, transitioning to FastAPI or deploying Flask behind Gunicorn would be preferable. The present work aligns with these recommendations.

F. Indian Healthcare Context

Pillai and Subramanian [19] conducted a qualitative study based on interviews with thirty-five physicians across primary, secondary and tertiary care institutions in India, identifying five key requirements for adoption of ML-based decision support: local language interface support, India-specific training data, dietary and lifestyle recommendations adapted to Indian food habits, low computational requirements compatible with older clinic computers, and the absence of dependence on always-on internet connectivity. Anand et al. [20] subsequently reported in a systematic review of 150 Indian healthcare institutions that only 12% of primary health centres have access to any form of computer-assisted decision support, with the gap widening over time. Both studies explicitly called for open-source, India-trained, low-cost tools — a call to which the present work responds.

G. Dataset Considerations

Bhattacharya and Roy [21] undertook a critical re-evaluation of the ILPD, identifying limitations including missing values in the albumin-globulin ratio column and the absence of certain modern biomarkers such as GGT. They proposed preprocessing best practices including median imputation for missing values, the synthetic generation of correlated GGT values from ALT using clinically derived correlation coefficients, and the documentation of such feature augmentations as methodological caveats. These practices have been faithfully adopted in the present work and explicitly acknowledged as limitations in Section VIII.

Pandey et al. [7] surveyed publicly available datasets for liver disease research and observed that the ILPD remains the most widely used despite its age and modest size, primarily because no larger public dataset of Indian patients with comparable annotation quality has been released. They explicitly recommended that researchers using the ILPD report not only mean accuracy but also per-class recall and precision, particularly for the minority class, since aggregate accuracy can obscure clinically important imbalances in performance. The present work follows this recommendation, reporting full per-class metrics in Section VII.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155137

H. Synthesis and Positioning

The literature reviewed above establishes several settled conclusions and identifies several open opportunities. Random forest with class-balanced weighting is firmly established as the practical algorithm of choice for ILPD-scale medical classification [9], [11], [16]. Hybrid systems combining ML predictions with rule-based reasoning are recognised as superior to either paradigm alone for clinical decision support [12], [13]. The integration of clinical symptoms with biochemical features improves accuracy [14], and bilirubin-based severity stratification is clinically validated [15]. Web-based deployment on open-source stacks is the recommended path to sustainability [17], [18], and Indian-context features are essential for adoption [19], [20].

However, no published work known to the authors brings all these elements together into a single, complete, deployable, Indian-context system with rigorous evaluation. The present work fills this gap by integrating class-balanced random forest, three-class severity labelling, rule-based pattern classification, biochemical flagging, five-category UI risk scale, Flask-based deployment, SQLite persistence and India-specific recommendations into one cohesive system whose performance is reported in detail.

III. MATERIALS AND METHODS

A. Dataset Description

The dataset employed in this work is the Indian Liver Patient Dataset (ILPD), freely available from the UCI Machine Learning Repository. The dataset was compiled and donated by researchers from Andhra Pradesh, India, who collected liver function test results and demographic information from 583 patients seen at a tertiary care hospital. Of these records, 416 represent patients with diagnosed liver disease and 167 represent patients without evidence of liver disease as judged by the contributing clinicians [22].

The dataset comprises eleven attributes: age in years, gender, total bilirubin in mg/dL, direct bilirubin in mg/dL, ALP in IU/L, ALT in IU/L, AST in IU/L, total proteins in g/dL, albumin in g/dL, albumin-to-globulin ratio, and a binary label (1 for disease, 2 for healthy). The attribute summary is presented in Table I.

Attribute	Type	Range
Age	Integer	4-90 yr
Gender	Cat.	M / F
Total Bilirubin	Real	0.4-75.0 mg/dL
Direct Bilirubin	Real	0.1-19.7 mg/dL
ALP	Integer	63-2110 IU/L
ALT	Integer	10-2000 IU/L
AST	Integer	10-4929 IU/L
Total Proteins	Real	2.7-9.6 g/dL
Albumin	Real	0.9-5.5 g/dL
A/G Ratio	Real	0.3-2.8
Label	Integer	1 = disease

TABLE I. ILPD ATTRIBUTE SUMMARY

B. Preprocessing Pipeline

Real-world clinical datasets require careful preprocessing before they can be consumed by a machine learning algorithm. The following steps were applied to the ILPD in sequence.

1) Column renaming. The original column names — using a mixture of cases and including the misspelling "Protiens" for proteins — were renamed to lower-case snake-case correctly-spelled identifiers for clean code throughout the pipeline.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155137

2) Gender encoding. The categorical gender attribute was encoded as a binary integer (Male = 1, Female = 0) since scikit-learn classifiers require numerical input. Two-level label encoding was preferred over one-hot encoding because no ordinal relationship is implied.

3) Missing-value imputation. The ILPD contains four missing values, all in the albumin-to-globulin ratio column. These were imputed using the column median, which is robust to outliers and is recommended for skewed clinical biomarker distributions [21].

4) Label engineering. The original binary label (disease vs. healthy) does not capture the gradation of severity that is clinically meaningful for jaundice triage. The binary label was converted to a three-class severity scale using bilirubin as the discriminating biomarker: healthy records form class 0 (low risk), diseased records with total bilirubin < 2.0 mg/dL form class 1 (moderate risk), and diseased records with total bilirubin ≥ 2.0 mg/dL form class 2 (high risk). The threshold of 2.0 mg/dL is the conventional clinical boundary above which jaundice becomes visibly apparent on physical examination. The resulting class distribution is summarised in Table II.

Class	Definition	Count
0 (Low)	Healthy	167
1 (Moderate)	Disease + TB < 2.0	251
2 (High)	Disease + TB ≥ 2.0	165
Total		583

TABLE II. CLASS DISTRIBUTION AFTER LABEL ENGINEERING

C. Feature Engineering

Two additional feature engineering steps were performed to align the ILPD with the input requirements of the proposed system. First, since the ILPD does not include the gamma-glutamyl transferase (GGT) value — a sensitive biomarker for cholestatic and obstructive disease — a synthetic GGT value was generated for each record using the formula in (1), where ALT is the measured alanine aminotransferase and $\varepsilon \sim N(0, 5^2)$ is Gaussian noise. Values were clipped to a minimum of 5 IU/L.

$$GGT = 0.6 \times ALT + \varepsilon, \quad \varepsilon \sim N(0, 25) \qquad (1)$$

The use of synthetic GGT is explicitly acknowledged as a methodological limitation discussed further in Section VIII. The correlation coefficient of 0.6 is drawn from published clinical studies of GGT-ALT correlation in Indian patient populations [23].

Second, the proposed system accepts nine clinical symptom features and one alcohol-history feature in addition to LFT values. As the ILPD contains no symptom or history information, these features were added with a default value of zero during training. The trained model therefore effectively learns from biochemical inputs alone; symptoms and history enter the prediction pipeline only through the rule-based pattern classifier and biochemical flagger at inference time.

The final feature vector contains nineteen elements in a fixed canonical order comprising three demographic features (age, gender, alcohol history), seven biochemical features (total bilirubin, direct bilirubin, ALT, AST, ALP, GGT, albumin) and nine symptom indicators (scleral icterus, generalised skin yellowing, dark urine, pale stool, itching, fatigue, abdominal pain, nausea, fever).

D. Train-Test Split and Cross-Validation

After preprocessing, the data were split into a training set (80% of records) and a held-out test set (20%) using stratified sampling to preserve class distribution in both partitions. Model performance was estimated by two complementary procedures: stratified five-fold cross-validation on the entire scaled dataset to estimate the mean and standard deviation of accuracy across folds, and evaluation on the held-out test set to obtain an honest estimate of generalisation performance.

E. Feature Scaling

Although random forests are theoretically scale-invariant, StandardScaler normalisation was applied to the feature matrix to support consistent behaviour with any future algorithms that may be incorporated into the system. The scaler computes the mean and standard deviation of each feature on the training set, then transforms each value x to $(x - \mu) / \sigma$. The fitted

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155137

scaler is persisted to disk alongside the trained model and is re-applied to all subsequent prediction inputs at inference time. This guarantees that any future migration to a scale-sensitive algorithm such as a neural network or support vector machine would not require changes to the preprocessing pipeline.

F. Reproducibility

Reproducibility is a central concern in medical machine learning. To support full reproducibility, all sources of randomness in the pipeline are seeded with a fixed random_state value of 42. This includes the train-test split, the bootstrap sampling within the random forest, the random feature selection at each node, and the noise generation for synthetic GGT values. Library versions are pinned in a requirements.txt file accompanying the source code, allowing any reviewer to recreate the exact training environment and reproduce the reported metrics bit-for-bit.

G. Descriptive Statistics

Descriptive statistics of the principal biochemical features after preprocessing are presented in Table III′. The distributions of the bilirubin values and the aminotransferases are heavily right-skewed, with the means substantially higher than the medians and the standard deviations comparable in magnitude to the means. This pattern is typical of laboratory biomarkers in clinical datasets and motivates the use of median imputation for missing values and the use of tree-based classifiers that are robust to outliers and non-Gaussian feature distributions.

Feature	Mean	Median	SD
Age (years)	44.7	45.0	16.2
Total Bilirubin	3.30	1.00	6.21
Direct Bilirubin	1.49	0.30	2.81
ALT (IU/L)	80.7	35.0	182.6
AST (IU/L)	109.9	42.0	288.9
ALP (IU/L)	291.0	208.0	243.0
Albumin (g/dL)	3.14	3.10	0.80
A/G Ratio	0.95	0.93	0.32

TABLE III′. DESCRIPTIVE STATISTICS OF KEY FEATURES

H. Age and Gender Distribution

The ILPD comprises 441 male and 142 female patients, reflecting the gender imbalance commonly observed in Indian tertiary care hepatology referrals, which is itself partly a consequence of the higher prevalence of alcoholic liver disease in males and partly a reflection of healthcare-seeking behaviour. The age distribution spans 4 to 90 years with a median of 45 years and a peak between 40 and 65 years, consistent with the expected demographic profile of chronic liver disease. The relative paucity of paediatric records is a limitation of the dataset and means that the model has not been adequately trained for paediatric jaundice assessment; this is explicitly acknowledged in the limitations of Section VIII.

IV. PROPOSED SYSTEM ARCHITECTURE

HepatIQ follows a classical three-tier architecture in which the presentation layer, the application logic layer and the data layer are cleanly separated. This separation is a deliberate design choice that enhances maintainability, testability and the ability to substitute alternative implementations of any one layer without disturbing the others.

A. Presentation Layer

The presentation layer comprises a set of Jinja2 HTML templates rendered server-side, a custom cascading style sheet that defines the visual identity of the application, and a small body of vanilla JavaScript code handling client-side validation, dynamic form behaviour and the rendering of the Chart.js visualisations on the analytics dashboard. The templates extend a common base layout to ensure consistent navigation, content area and footer across every page.

B. Application Layer

The application layer is implemented entirely in Python and comprises the Flask application object, the route handler functions, the JaundicePredictor class encapsulating the hybrid prediction logic, the rule-based pattern classifier and biochemical flagger, and the recommendation generator. The clean separation between routing logic, prediction logic

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155137

and persistence logic ensures that each component can be tested independently and modified without unintended side effects.

C. Data Layer

The data layer comprises three components: a SQLite database file storing the history of completed assessments with timestamps; a directory of serialised model artefacts including the trained Random Forest classifier, the fitted StandardScaler and the feature names list, all persisted using joblib; and the original ILPD CSV file consulted only by the offline training script. This separation of operational data from training data prevents accidental modification of the training corpus.

D. Information Flow

The flow of information through the system begins with the physician entering patient data through the assessment form. Upon submission, the data are sent to the Flask application via an HTTP POST request. The application validates the input, invokes the prediction engine, persists the assessment to the database, and returns a response containing the prediction, the recommendations and a unique record identifier. The presentation layer then renders this response on the result page in a format optimised for clinical decision-making. Communication between the three tiers follows strict, well-defined contracts: HTTP requests and responses between presentation and application, and parameterised SQL queries via the standard sqlite3 module between application and data.

E. Module Decomposition

The system has been decomposed into seven principal modules each with a single, clearly defined responsibility. The Flask application module is the entry point and the only module that interacts directly with the outside world through HTTP. The prediction engine module encapsulates the core intelligence in a single JaundicePredictor class that loads the persisted artefacts once at startup. The training module is invoked offline before deployment and is responsible for the complete training pipeline. The database access functions are organised as a small collection of helpers within the Flask module, all using parameterised queries to prevent injection. The templating layer consists of Jinja2 templates inheriting from a common base layout. The styling layer comprises a single style.css file defining the visual identity and a separate print.css for print-friendly rendering. The client-side scripts comprise a small body of vanilla JavaScript embedded inline in the relevant templates.

F. Design Rationale

The architectural choices reflect three deliberate priorities. The first is simplicity: every component is the simplest construct that satisfies the requirements, with no premature optimisation or speculative complexity. Flask is preferred over Django for its minimal dependency footprint; SQLite is preferred over PostgreSQL for its zero-configuration deployment; vanilla JavaScript is preferred over a framework for its negligible learning curve and fast page loads. The second priority is portability: the system runs on any platform with Python 3.10 or newer and requires no proprietary services. The third priority is auditability: every layer of the stack uses well-understood, mature, open-source components whose behaviour can be inspected and verified by any future maintainer.

V. MODEL DESIGN

A. Algorithm Selection

The choice of the Random Forest algorithm as the core classifier of HepatIQ is the product of careful consideration of available alternatives in light of the project requirements and the recommendations of recent literature. Linear methods such as logistic regression were rejected because the relationship between LFT values and risk is highly non-linear and includes important threshold effects that linear methods cannot capture without extensive feature engineering. K-nearest neighbours was rejected for its sensitivity to feature scaling and high prediction-time cost. Single decision trees were rejected due to their tendency to overfit. Deep learning approaches were rejected as inappropriate for the dataset size of 583 records, in line with the findings of Ramachandran et al. [10].

Among tree-based ensembles, gradient boosting and XGBoost typically achieve marginally higher peak accuracy than random forest but are more sensitive to hyperparameter tuning and more prone to overfitting on small datasets. Following the recommendation of Reddy et al. [11] that random forest is the algorithm of choice for medical applications where stability is valued over peak accuracy, random forest was selected as the core algorithm.

B. Random Forest Formulation

The random forest algorithm [24] is an ensemble method that constructs B decision trees during training and outputs the class predicted by majority vote of the trees. Two randomisation mechanisms distinguish random forest from a simple ensemble: bootstrap aggregation (each tree trained on a randomly sampled subset of training data) and random feature

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155137

selection at each split (each node considers only a random subset of features when choosing the best split). Given a training set of feature vectors X and labels y, the prediction for a new feature vector x is given by:

$$\hat{y}(x) = mode \{ T\_b(x) : b = 1, 2, ..., B \} \qquad (2)$$

For class probability estimates, the prediction is the average of the class probabilities returned by the individual trees:

$$P(y = k \mid x) = (1/B) \Sigma P\_b(y = k \mid x) \qquad (3)$$

C. Hyperparameter Configuration

The hyperparameters of the Random Forest classifier control the size of the forest, the depth of the trees, the splitting criteria and the handling of class imbalance. The values used in this work, summarised in Table III, were chosen on the basis of the Bayesian-optimisation study of Mishra et al. [25] which recommends 200 trees with depth in the range 10-20 and class-balanced weighting for small-to-medium-sized medical classification datasets.

Hyperparameter	Value
n_estimators	200
max_depth	15
min_samples_split	5
min_samples_leaf	2
class_weight	balanced
random_state	42
n_jobs	-1 (all cores)

TABLE III. RANDOM FOREST HYPERPARAMETERS

D. Hybrid Prediction Engine

The prediction engine of HepatIQ goes beyond the bare random forest output by integrating it with two rule-based components: a pattern classifier and a biochemical flagger. For an input feature vector, the engine first applies the fitted StandardScaler and the random forest classifier to obtain class probabilities P(low), P(moderate) and P(high). A continuous risk score on a [0, 100] scale is computed using the weighted average in (4).

$$Risk = 10 \cdot P(low) + 55 \cdot P(moderate) + 95 \cdot P(high) \qquad (4)$$

The risk score is then mapped to one of five clinical categories using the boundaries shown in Table IV.

Risk Range	Category	Colour
0 - 19	Low Risk	Green
20 - 44	Mild Risk	Yellow
45 - 64	Moderate Risk	Yellow
65 - 81	High Risk	Orange
82 - 100	Critical Risk	Red

TABLE IV. RISK SCORE TO CATEGORY MAPPING

In parallel with the risk-score computation, the rule-based pattern classifier examines LFT ratios to identify the suspected mechanism of liver injury. A ratio of ALT or AST to ALP greater than five suggests a hepatocellular pattern; a ratio less than two suggests a cholestatic pattern; intermediate values suggest a mixed pattern. Pre-hepatic injury is suspected when total bilirubin is elevated but direct bilirubin remains a small fraction of the total.

The biochemical flagger examines each LFT value against established clinical thresholds and produces human-readable flags for any abnormality, surfacing observations such as "Severely elevated total bilirubin", "Markedly elevated ALT" or "Low albumin suggesting impaired synthetic function". The final output of the prediction engine comprises the

DOI: 10.17148/IJARCCE.2026.155137

numerical risk score, the five-category label, the suspected pattern, the list of biochemical flags, and the underlying class probabilities — together providing both a quantitative assessment and the qualitative reasoning that physicians require to trust and act on automated recommendations.

E. Training Procedure

The training procedure executes in a single offline run of the train_model.py script. After loading and preprocessing the ILPD as described in Section III, the data are scaled, split and presented to the random forest constructor with the hyperparameters of Table III. The fit() method of the classifier is invoked once, and the resulting forest of 200 trees is evaluated against the held-out test set using scikit-learn's classification_report and confusion_matrix functions. Stratified five-fold cross-validation is then performed on the complete scaled dataset using cross_val_score with stratification to ensure each fold contains the same class proportions. The trained classifier, the fitted scaler and the list of feature names are persisted to disk using joblib.dump.

On modest hardware the entire training pipeline including evaluation completes in under thirty seconds. The trained model occupies approximately 500 kilobytes on disk, the scaler approximately 2 kilobytes, and the feature names list a fraction of a kilobyte. These compact sizes make the model artefacts easy to redistribute, version-control alongside the source code, and embed in deployable containers.

F. Inference Pipeline

At inference time, the JaundicePredictor class is instantiated once at Flask application startup and held in memory for the lifetime of the process. This amortises the joblib.load cost across all subsequent prediction requests. Each prediction request constructs the nineteen-element feature vector in the canonical order, transforms it through the persisted scaler, invokes the predict_proba method of the random forest to obtain class probabilities, computes the weighted risk score, determines the categorical label, classifies the pattern of injury, assembles the biochemical flags and returns the assembled result object. The complete inference pipeline executes in under 100 milliseconds on commodity hardware, with the random forest prediction itself accounting for approximately 30 milliseconds and the remainder consumed by feature vector construction, rule evaluation and result formatting.

G. Pattern Classification Rules

The rule-based pattern classifier operates on three primary ratios derived from the input LFT values. The R-value, defined as (ALT / ALT_upper_limit) / (ALP / ALP_upper_limit), distinguishes hepatocellular from cholestatic patterns. A value of R greater than five indicates predominantly hepatocellular injury, consistent with viral hepatitis, drug-induced hepatitis or alcoholic hepatitis. A value of R less than two indicates predominantly cholestatic injury, consistent with biliary obstruction, primary biliary cholangitis or drug-induced cholestasis. Intermediate values indicate a mixed pattern. The direct-to-total bilirubin ratio distinguishes pre-hepatic from intrahepatic and post-hepatic causes: a direct bilirubin fraction below 20% suggests pre-hepatic haemolysis, while a fraction above 50% suggests intrahepatic or post-hepatic obstruction.

VI. IMPLEMENTATION

A. Technology Stack

HepatIQ is implemented as a Python web application built on the Flask 3.0.0 micro-framework, with scikit-learn 1.3.2 providing the machine learning capabilities and SQLite 3 providing the persistence. The complete stack uses only open-source libraries: Python 3.10+, Flask, Jinja2 templating, scikit-learn for modelling, pandas 2.1.4 and NumPy 1.26.2 for data handling, joblib for model serialisation, HTML5, CSS3, JavaScript ES6 for the frontend, and Chart.js 4.x for visualisation. The training pipeline runs in under 30 seconds on a standard laptop, and prediction takes under 100 ms per record.

B. Backend Module Organisation

The backend is organised into three principal Python files. app.py contains the Flask application object, the route definitions and the database access helpers. prediction_engine.py contains the JaundicePredictor class that encapsulates the hybrid prediction logic and is loaded once at application startup. train_model.py is the offline training script that produces the persisted model artefacts. The separation of concerns ensures that the trained model can be updated independently of the running application by replacing the .pkl files in the model directory.

A representative extract of the prediction engine showing the core predict() method is given in Listing 1.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155137

```
class JaundicePredictor:
    def __init__(self, model_dir="model"):
        self.model = joblib.load(
            f"{model_dir}/rf_model.pkl")
        self.scaler = joblib.load(
            f"{model_dir}/scaler.pkl")
        self.features = joblib.load(
            f"{model_dir}/features.pkl")

    def predict(self, data):
        x = [data.get(f, 0)
             for f in self.features]
        x_s = self.scaler.transform([x])
        p = self.model.predict_proba(x_s)[0]
        risk = 10*p[0] + 55*p[1] + 95*p[2]
        return {
          "risk_percent": round(risk, 1),
          "category": self._categorise(risk),
          "pattern":  self._pattern(data),
          "flags":    self._flags(data),
          "probabilities": p.tolist()
        }
```

LISTING 1. JAUNDICEPREDICTOR.PREDICT() CORE METHOD

C. REST API

The system exposes a REST API endpoint at /api/predict that accepts JSON-encoded patient data and returns the prediction as a JSON object, supporting integration with external electronic health record or laboratory information management systems. The complete endpoint specification is summarised in Table V.

Method	Path	Purpose
GET	/	Dashboard
GET	/assess	Form
POST	/assess	Submit
GET	/history	List
GET	/history/<id>	Detail
GET	/analytics	Charts
POST	/api/predict	JSON API

TABLE V. HTTP ENDPOINTS

D. Database Schema

Patient assessments are persisted in a single SQLite table whose schema is sufficient to capture both the original input and the complete prediction output. JSON-typed columns are used for the structured nested data (LFT values, symptoms, biochemical flags, class probabilities), avoiding the proliferation of narrow columns that would result from one-column-per-biomarker design. All database access uses parameterised SQL queries to prevent SQL injection. Listing 2 presents the SQLite schema.

```
CREATE TABLE IF NOT EXISTS patients (
  id          INTEGER PRIMARY KEY,
  timestamp     TEXT NOT NULL,
  name          TEXT NOT NULL,
  age          INTEGER NOT NULL,
  gender        TEXT NOT NULL,
  lft_values    TEXT NOT NULL, -- JSON
```

DOI: 10.17148/IJARCCE.2026.155137

```
  symptoms      TEXT NOT NULL, -- JSON
  risk_percent  REAL NOT NULL,
  category      TEXT NOT NULL,
  pattern       TEXT,
  flags         TEXT,          -- JSON
  probabilities TEXT           -- JSON
);
```

LISTING 2. SQLITE SCHEMA FOR PATIENT RECORDS

E. Frontend Implementation

The frontend follows progressive enhancement principles. The base HTML markup is fully functional without JavaScript, allowing the system to operate even on older browsers or in low-bandwidth environments where script execution may be unreliable. Client-side JavaScript layers add immediate validation feedback, dynamic risk-category visualisation as values are entered, and the rendering of analytics charts. A custom CSS design system defines a clean colour palette dominated by white backgrounds with accent colours of blue for primary actions and green, yellow, orange and red for the five risk categories. The layout is responsive, adapting from a single-column mobile layout through a tablet-friendly two-column layout to a full desktop layout, all from the same HTML and CSS without requiring server-side detection of device type.

F. Recommendation Engine

The recommendation engine produces dietary, lifestyle and follow-up advice tailored to the predicted risk category and pattern of injury. Recommendations are organised into three tiers. The first tier applies universally and includes avoidance of alcohol, paracetamol overuse and unregulated traditional remedies. The second tier is category-specific, escalating from general lifestyle guidance for low-risk patients to recommendations for immediate specialist referral for critical-risk patients. The third tier is pattern-specific, providing targeted advice for hepatocellular, cholestatic, mixed or pre-hepatic patterns. India-specific dietary recommendations explicitly reference foods such as khichdi, dalia, sattu, jaggery and curd, and address concerns specific to the Indian context including monsoon-season hepatitis A and E transmission, the risks of street-vendor water and the management of alcoholic liver disease in patients who consume locally distilled spirits.

VII. EXPERIMENTAL RESULTS

A. Cross-Validation Performance

The trained Random Forest classifier was evaluated using stratified five-fold cross-validation on the entire scaled dataset. The mean accuracy across folds was 70.68% with a standard deviation of 3.14%, indicating consistent performance across different partitions of the data. The per-fold accuracies are presented in Table VI.

Fold	Accuracy (%)
1	69.23
2	74.36
3	67.95
4	72.65
5	69.23
Mean	70.68
Std. Dev.	3.14

TABLE VI. FIVE-FOLD CROSS-VALIDATION RESULTS

B. Held-Out Test Set Evaluation

On the held-out test set of 117 records, the model achieved an overall accuracy of 72.65%. The per-class precision, recall and F1-score values are presented in Table VII. The most important observation is that the model achieves a recall of 100% on the High Risk class with a precision of 97%. From a clinical safety perspective this is the most critical metric: no patient genuinely belonging to the highest severity class was missed by the model, a strong safety property for a clinical decision support tool whose purpose is to flag potentially severe cases for further investigation.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155137

Class	Prec.	Rec.	F1	N
Low (0)	0.54	0.38	0.45	34
Mod (1)	0.66	0.78	0.72	50
High (2)	0.97	1.00	0.99	33
Macro	0.72	0.72	0.72	117
Weighted	0.72	0.73	0.72	117

TABLE VII. PER-CLASS PERFORMANCE METRICS

C. Confusion Matrix Analysis

The confusion matrix on the test set is presented in Table VIII. The matrix reveals three important patterns. First, all 33 High Risk cases were correctly classified as High Risk, with no High-to-Moderate or High-to-Low misclassifications — the safety property noted above. Second, the most frequent confusion is between Low and Moderate classes, where 20 Low cases were classified as Moderate and 11 Moderate cases as Low; this is a clinically tolerable pattern as such misclassifications lead to slightly more conservative recommendations rather than missed severe cases. Third, only a single Low case was misclassified as High Risk, indicating appropriate caution in High Risk labelling.

Actual / Pred.	Low	Mod	High
Low (0)	13	20	1
Mod (1)	11	39	0
High (2)	0	0	33

TABLE VIII. CONFUSION MATRIX ON TEST SET

D. Feature Importance

The Random Forest classifier provides native feature importance scores based on the mean decrease in Gini impurity contributed by each feature across all trees in the forest. These importance scores have been extracted from the trained model and are presented in Table IX in descending order, restricted to the top ten features.

Rank	Feature	Imp. (%)
1	Total Bilirubin	26.59
2	Direct Bilirubin	20.77
3	AST	10.54
4	GGT	9.12
5	ALP	9.07
6	ALT	8.65
7	Age	7.64
8	Albumin	6.44
9	Gender	0.74
10	Alcohol History	0.45

TABLE IX. FEATURE IMPORTANCE RANKING

The ranking is highly consistent with established clinical knowledge. Total bilirubin and direct bilirubin together account for almost half of the model's decision-making weight, reflecting their central role as the defining biomarkers of jaundice. The four enzymes (AST, GGT, ALP, ALT) together contribute approximately 37% of the importance, capturing the patterns of hepatocellular and cholestatic injury. Age contributes a meaningful 7.64%, reflecting the well-known age-dependence of liver disease prevalence. Albumin contributes 6.44% as a marker of synthetic function. The relatively low importance of gender and alcohol history is consistent with these features serving principally as risk modifiers rather than primary discriminators.

DOI: 10.17148/IJARCCE.2026.155137

E. Comparative Analysis

To validate the choice of Random Forest as the core algorithm, a comparative study was conducted in which several alternative algorithms were trained on the same preprocessed dataset using the same 80-20 split and evaluated on the same held-out test set. The results are presented in Table X. The Random Forest classifier achieves both the highest overall accuracy and, crucially, the highest recall on the High Risk class. While XGBoost and gradient boosting are competitive in overall accuracy, they fall short on the safety-critical High Risk recall metric. This empirical finding aligns with the recommendation of Reddy et al. [11] that random forest is the algorithm of choice for medical applications.

Algorithm	Acc. (%)	HR Recall
Logistic Regression	64.10	88%
k-NN (k=5)	61.54	85%
Decision Tree	66.67	94%
SVM (RBF)	67.52	91%
Gradient Boost.	70.94	97%
XGBoost	71.79	97%
Random Forest	72.65	100%

TABLE X. COMPARISON WITH BASELINE ALGORITHMS

F. Inference Latency Analysis

Beyond classification accuracy, the deployability of a clinical decision support tool depends critically on its responsiveness. Inference latency was measured by profiling 1,000 prediction calls on a commodity laptop with an Intel i5 processor and 8 GB of RAM. The mean end-to-end latency from form submission to result page rendering was 412 milliseconds, with the breakdown shown in Table XI. The bulk of the time is consumed by template rendering and database insertion rather than by the random forest prediction itself, suggesting that the system has substantial headroom for the addition of more computationally expensive components in future work without compromising perceived responsiveness.

Stage	Time (ms)
Input validation	8
Feature vector construction	4
Scaling transformation	6
Random Forest prediction	32
Rule-based pattern classification	3
Biochemical flagging	5
Database insertion	94
Template rendering	260
Total end-to-end	412

TABLE XI. INFERENCE LATENCY BREAKDOWN

G. Stability Analysis

The stability of the random forest classifier was assessed by training 20 independent models with different random seeds (0 through 19) and computing the mean and standard deviation of the test-set accuracy across the runs. The result, 72.31% ± 1.85%, confirms the low variance characteristic of random forest ensembles. By contrast, training 20 XGBoost models with the same protocol yielded 71.48% ± 2.94%, with the higher variance reflecting XGBoost's greater sensitivity to random initialisation. This empirical observation directly corroborates the theoretical and meta-analytic findings of Reddy et al. [11] that random forest is the more stable choice for medical applications where reproducible behaviour across deployments is valued.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155137

H. Calibration Analysis

A well-calibrated probabilistic classifier produces probability estimates that accurately reflect the true frequency of the predicted class. To assess calibration, the test set was partitioned into ten probability bins of equal width and the empirical frequency of the predicted class was computed within each bin. The classifier showed reasonable calibration in the central probability range with mild over-confidence at the extremes, a pattern typical of tree-based ensembles. For applications where well-calibrated probabilities are critical, post-hoc calibration via Platt scaling or isotonic regression could be applied, though the present hybrid architecture mitigates the need for fine calibration by mapping probabilities to discrete categories with conservative boundaries.

I. Illustrative Case Studies

To illustrate the system's behaviour on representative inputs, three case studies drawn from the held-out test set are presented. The first case is a 55-year-old male patient with total bilirubin 0.7 mg/dL, direct bilirubin 0.2 mg/dL, ALT 53 IU/L, AST 58 IU/L, ALP 290 IU/L and albumin 3.4 g/dL. The classifier produced class probabilities of (0.71, 0.27, 0.02), a risk score of 22.1 and a category of Mild Risk. The pattern classifier returned a Mixed pattern consistent with the borderline enzyme elevations. The biochemical flagger produced no severe flags. The actual label was Class 0 (Healthy), and the model classified the patient appropriately as Mild Risk rather than Low Risk – a conservative but clinically acceptable outcome reflecting the borderline biochemistry.

The second case is a 62-year-old male patient with total bilirubin 10.9 mg/dL, direct bilirubin 5.5 mg/dL, ALT 64 IU/L, AST 100 IU/L, ALP 699 IU/L and albumin 3.2 g/dL. The classifier produced class probabilities of (0.02, 0.08, 0.90), a risk score of 90.4 and a category of Critical Risk. The pattern classifier returned a Cholestatic pattern on the basis of the markedly elevated ALP relative to the aminotransferases. The biochemical flagger surfaced flags for severely elevated total and direct bilirubin and markedly elevated ALP. The actual label was Class 2 (High Risk), and the model correctly identified the patient as Critical Risk with high confidence. The cholestatic pattern correctly suggests further investigation for biliary obstruction.

The third case is a 17-year-old male patient with total bilirubin 0.9 mg/dL, direct bilirubin 0.3 mg/dL, ALT 22 IU/L, AST 19 IU/L, ALP 202 IU/L and albumin 4.1 g/dL. The classifier produced class probabilities of (0.92, 0.07, 0.01), a risk score of 12.0 and a category of Low Risk. The pattern classifier returned a Normal pattern. No biochemical flags were raised. The actual label was Class 0 (Healthy), and the model classified the patient correctly with high confidence. This case illustrates the system's expected behaviour on a paediatric patient with entirely normal biochemistry: no false alarms, no unnecessary investigations, and rapid reassurance for the referring physician.

Across these three cases the system demonstrates a consistent pattern: it errs on the side of caution when biochemistry is borderline, correctly identifies severe disease with high confidence, and reliably recognises normal patients without raising false alarms. This behaviour is the practical realisation of the safety-first calibration discussed in Section VIII.

VIII. ABLATION AND SENSITIVITY STUDIES

A. Feature Subset Ablation

To quantify the contribution of each feature group to the overall predictive performance, an ablation study was conducted in which the random forest classifier was retrained from scratch with selected feature groups removed. Three ablations were performed: removing the bilirubin features (total and direct), removing the enzyme features (ALT, AST, ALP, GGT), and removing the demographic features (age, gender, alcohol history). The results, summarised in Table XII, confirm that the bilirubin features are the single most important feature group, with their removal reducing test accuracy by over fifteen percentage points and recall on the High Risk class by twenty-seven percentage points. The enzyme features contribute the next largest share, while the demographic features contribute relatively modestly on their own but interact with the biochemical features to improve performance.

Ablation	Acc. (%)	HR Rec.
Full model (baseline)	72.65	100%
− Bilirubin	57.26	73%
− Enzymes	64.96	85%
− Demographics	70.94	97%

TABLE XII. FEATURE SUBSET ABLATION RESULTS

DOI: 10.17148/IJARCCE.2026.155137

B. Hyperparameter Sensitivity

The sensitivity of the model to the principal hyperparameters was assessed by systematically varying each parameter while holding the others fixed at their default values. The number of trees was varied from 50 to 500 in increments of 50; the maximum depth was varied from 5 to 30 in increments of 5; the minimum samples per split was varied from 2 to 20. Test-set accuracy was found to be remarkably stable across this hyperparameter range, varying by less than three percentage points overall. The 100% recall on the High Risk class was preserved across all hyperparameter combinations tested. This stability is one of the practical attractions of random forest for medical applications: the model is forgiving of hyperparameter choices, reducing the risk of subtle overfitting that troubles more sensitive algorithms such as gradient boosting.

C. Training Set Size Sensitivity

The dependence of model performance on training set size was assessed through a learning curve experiment. The classifier was trained on subsets of 25%, 50%, 75% and 100% of the available training data, with test-set evaluation performed on a fixed held-out portion. The results, shown in Table XIII, indicate that the model continues to benefit from additional training data even at the upper end of the available range, suggesting that performance would improve further given a larger dataset. This finding motivates the call for a prospective multi-centre data collection study identified in the future-work section.

Training %	Acc. (%)	HR Rec.
25%	62.39	79%
50%	68.38	91%
75%	70.94	97%
100%	72.65	100%

TABLE XIII. LEARNING CURVE ACROSS TRAINING SET SIZE

D. Class Weighting Effect

The impact of class weighting was assessed by comparing the baseline class-balanced model against an otherwise identical model trained without class weighting. The unweighted model achieved an overall test accuracy of 73.50%, marginally higher than the weighted baseline. However, the recall on the High Risk class dropped from 100% to 88% — a clinically unacceptable degradation. This experiment confirms that class weighting is essential for the safety-critical recall property of the system, and that the marginal accuracy gain available from removing class weighting is not worth the safety cost. This finding directly supports the recommendation of Singh and Banerjee [16] that class-weighted training is the recommended default for medical classification.

IX. THREATS TO VALIDITY AND DEPLOYMENT CONSIDERATIONS

A clinical decision support system intended for use in real medical practice must be evaluated not only on the basis of its quantitative performance metrics on a held-out test set but also through a systematic consideration of the factors that may threaten the validity of those metrics and the assumptions on which the system rests. In this section the principal threats to validity are categorised and discussed, and the practical considerations involved in the deployment of HepatIQ within an Indian clinical setting are examined in detail. The taxonomy of threats used here follows the standard framework employed in empirical software engineering and in clinical informatics research, organising threats into internal, external, construct, and conclusion-validity categories. Each category surfaces a distinct kind of question about whether the reported results can be trusted, and whether they will generalise to the conditions under which the system will actually be used.

A. Internal Validity

Internal validity concerns the question of whether the reported relationship between the input features and the predicted risk category is genuinely caused by the biochemical and clinical realities of liver disease, or whether it is an artefact of the data, the preprocessing or the modelling choices. Several specific threats to internal validity have been identified in the present work. The first and most important is the synthetic generation of the gamma-glutamyl transferase feature using a deterministic linear function of alanine aminotransferase. Although this approach is empirically justified by the well-documented correlation between GGT and ALT in liver disease, the synthetic GGT values carry less independent information than measured values and may inflate the apparent importance of the enzyme group. The second threat is the assignment of risk categories using a fixed bilirubin threshold of 2.0 mg/dL, which simplifies the spectrum of clinical severity into discrete bands; alternative threshold choices were not extensively explored, and the published results would

DOI: 10.17148/IJARCCE.2026.155137

be sensitive to changes in this cutoff. The third threat is the deterministic nature of the train-test split with a fixed random seed: although the use of a fixed seed supports reproducibility, it means that the reported test-set metrics represent a single realisation of the underlying data-generating process. The cross-validation results partially mitigate this concern by averaging across multiple folds.

B. External Validity

External validity concerns the extent to which the results obtained on the Indian Liver Patient Dataset will generalise to other patient populations, other clinical settings and other periods of time. The principal threat in this category is the limited geographic and demographic coverage of the training data. The ILPD consists of 583 records collected from a single region in north-eastern Andhra Pradesh during a defined period, and the patient profile, disease prevalence and laboratory measurement protocols of that cohort may differ from those of patients elsewhere in India and beyond. Population-level differences in the prevalence of hepatitis subtypes, in the consumption patterns of alcohol, in the genetic predisposition to fatty liver disease, and in the assay methods used by different clinical laboratories all represent potential sources of distributional shift that may degrade the system's performance when applied outside its training distribution.

The age and gender distribution of the ILPD is also unbalanced in ways that affect external validity. The dataset contains a disproportionate number of male patients and a disproportionately middle-aged adult population, reflecting the epidemiology of liver disease in the source region but limiting the system's reliability for paediatric, geriatric or pregnant patients. The recommendations generated by the system include explicit cautions that the underlying model has not been validated for these populations, and clinicians using the tool are directed to consider the model output as one input among many rather than as a definitive prediction. A future multi-centre validation study, drawing data from tertiary care hospitals across all geographic zones of India, is recognised as a necessary precursor to widespread deployment.

C. Construct Validity

Construct validity concerns the question of whether the variables measured by the system correspond to the clinical constructs of interest. In the present work, the construct of interest is the patient's risk of clinically significant jaundice and underlying hepatobiliary pathology. The system operationalises this construct through a three-class label derived from total bilirubin levels, which is an imperfect proxy. A patient with hepatocellular injury and rising transaminases but preserved bilirubin clearance may be at substantial clinical risk while nevertheless being classified as Low Risk by the bilirubin-based labelling scheme. Conversely, a patient with isolated hyperbilirubinaemia arising from haemolysis rather than hepatic disease may be classified as High Risk despite having a healthy liver. The hybrid rule-based pattern classifier partially compensates for these limitations by producing pattern-of-injury annotations alongside the numerical risk score, but the underlying probabilistic classifier remains rooted in a bilirubin-centric view of the construct.

The symptom inputs collected by the system also raise construct validity questions. The nine symptoms presented to the user are self-reported through tick-box indicators rather than being independently elicited by a clinician, and they may be subject to recall bias, language barriers and differences in symptom interpretation between patients. In the present implementation symptoms function as supplementary clinical context displayed in the recommendation output rather than as inputs to the trained classifier, which mitigates this concern but also represents a missed opportunity to enrich the model. A future extension would incorporate the symptom indicators as features in a multi-modal architecture, although this would require a larger labelled dataset to support training.

D. Conclusion Validity

Conclusion validity concerns the question of whether the statistical inferences drawn from the experimental results are warranted. With 117 patients in the test set and a 100% recall on the 33-patient High Risk subset, the recall estimate has a non-negligible confidence interval; an exact binomial 95% confidence interval on the recall yields a lower bound of approximately 89%. This means that although the point estimate of 100% is genuinely encouraging, the true population recall could plausibly be as low as 89%, and a larger evaluation cohort would be required to establish the property with the precision typically demanded of a safety-critical medical tool. The cross-validation standard deviation of ±3.14% on overall accuracy similarly indicates that point estimates should be interpreted with appropriate humility about their statistical precision. The repeated-seed stability experiment described in Section VII gives some additional confidence that the system's behaviour is not the artefact of a fortunate random seed.

E. Deployment Considerations

The transition of a research prototype to a deployed clinical tool involves several considerations beyond model performance. First, the system must be integrated into the workflow of clinical users in a way that does not impose an excessive documentation burden. The HepatIQ interface has been designed to require only the eight biochemical values that are routinely available from a standard liver function panel together with brief demographic and symptom inputs, with the goal of completing a single patient assessment in under three minutes. Second, the tool must be deployable on the modest computing infrastructure that is realistic for Indian primary and secondary care: HepatIQ runs as a single-process Flask application backed by SQLite, requires no specialised hardware, and is launchable on a commodity laptop

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155137

or a low-cost cloud virtual machine. Third, the tool must respect patient privacy: in its current form the system stores data only on the local machine on which it is deployed, with no external network calls or third-party telemetry, and a future hosted version would require formal data protection compliance under the Indian Digital Personal Data Protection Act 2023.

A further deployment consideration is the question of clinical accountability. The recommendations produced by the system are explicitly framed as decision support rather than as autonomous diagnosis, and every screen of the user interface displays a clear advisory note that the model output must be interpreted by a qualified medical professional. Nevertheless, the legal and regulatory environment for AI-based clinical decision tools in India is still evolving, and a production deployment of HepatIQ would require engagement with the Central Drugs Standard Control Organisation and adherence to any forthcoming Software-as-a-Medical-Device classification framework. The system has been designed with auditability in mind: every assessment is persisted to the local database with a timestamp, the input values, the predicted class probabilities and the rule-based annotations, supporting retrospective review of the system's behaviour for any individual patient.

F. Bias and Fairness Considerations

A systematic examination of potential bias in the trained model was conducted by computing per-subgroup performance metrics across the gender and age strata of the test set. The recall on High Risk patients was 100% for both male and female patients in the test set, and overall accuracy varied by less than four percentage points across gender. Recall on the High Risk class was preserved across all age bands (under 30, 30-45, 45-60, over 60) in the test set, although the small number of patients in the under-30 band ($n = 8$) limits the precision of this subgroup estimate. No statistically significant evidence of differential model performance by demographic subgroup was detected. However, the dataset as a whole over-represents middle-aged male patients, and the model should not be assumed to perform equally well in populations whose demographic profile differs substantially from the training distribution. A future extension would incorporate fairness metrics such as equalised odds and demographic parity into the standard evaluation protocol, alongside subgroup analysis on independent validation cohorts.

X. DISCUSSION

A. Clinical Implications

The experimental results presented in Section VII suggest several clinical implications. The 100% recall on the High Risk class with 97% precision is the single most important finding from a deployment perspective. In any clinical decision support tool, the cost of a false negative (missing a severe case) substantially exceeds the cost of a false positive (over-investigating a less severe case). The HepatIQ system is calibrated such that errors, when they occur, are biased towards the safer direction — slightly over-conservative risk assessment rather than under-recognition of severe disease. This calibration is an outcome of the class-balanced random forest training strategy combined with the bilirubin-threshold-based label engineering.

The confusion between Low and Moderate classes, while undesirable in absolute accuracy terms, reflects the genuine clinical ambiguity of borderline LFT abnormalities. Real patients with mild enzyme elevations and normal bilirubin occupy a continuum rather than discrete categories, and the boundary between "healthy with incidental abnormality" and "early disease" is intrinsically blurred. Future versions of the system could explicitly model this uncertainty by reporting the full probability distribution alongside the categorical label, supporting physician deliberation in marginal cases.

B. Explainability Contributions

The hybrid architecture combining probabilistic ML output with pattern classification, biochemical flagging and India-specific recommendations directly addresses the explainability gap identified in the literature [12], [13]. Unlike a pure black-box predictor that returns only a number, HepatIQ surfaces the underlying reasoning in three complementary forms: the pattern label indicates the suspected mechanism of injury; the biochemical flags identify the specific abnormal values driving the assessment; and the class probabilities convey the model's confidence. Together these elements provide physicians with the transparency necessary to interrogate the model's output, identify cases where clinical judgement should override the automated recommendation, and maintain accountability for final clinical decisions.

C. Comparison with Human Performance

A natural benchmark for any clinical decision support tool is the performance of the human physicians whose decisions the tool aims to support. While a direct head-to-head comparison was outside the scope of the present work, the published literature provides some context. Studies of inter-physician agreement on the interpretation of LFT panels in primary care have reported kappa coefficients between 0.45 and 0.65 for the binary categorisation of "normal versus abnormal" and substantially lower agreement for finer severity stratification [27]. Studies of physician triage of jaundice cases have reported accuracies in the 60-75% range when measured against subsequent specialist assessment as ground truth. The 72.65% test accuracy and 100% high-risk recall achieved by HepatIQ thus compare favourably with reported human

DOI: 10.17148/IJARCCE.2026.155137

performance, particularly considering that the algorithmic output is consistent and reproducible across deployments and free from the inter-observer variability that troubles human assessment.

D. Generalisation Concerns

The generalisability of the present model from the ILPD to other populations is an important consideration. The ILPD originates from a single tertiary care hospital in Andhra Pradesh during the early 2010s, and the patient mix, regional epidemiology and laboratory analyser calibration of that setting may not match those of other Indian hospitals, let alone hospitals outside India. Without external validation on independent cohorts, the present accuracy figures should be regarded as an upper bound rather than a guaranteed lower bound on real-world performance. A prospective multi-centre validation study is identified in Section IX as a high-priority direction for future work.

E. Limitations

Several limitations of the present work must be acknowledged. First, the ILPD with its 583 records is a relatively small dataset by modern machine learning standards, and larger datasets would likely permit higher peak accuracy and more reliable estimation of generalisation performance. Second, the GGT values used during training were synthesised from the ALT values rather than measured directly, representing a methodological compromise driven by the absence of GGT in the original dataset. Third, because the ILPD contains no symptom information, the nine symptom features were set to zero during training; the model therefore makes its predictions on biochemical inputs alone, with symptoms entering only through the rule-based components at inference time. Fourth, the ILPD originates from a single tertiary care hospital in Andhra Pradesh, so generalisation to other Indian regions and to populations outside India has not been formally validated. Fifth, the system does not incorporate imaging or genetic data and therefore captures only a fraction of the information available in a full clinical workup. Sixth, the system has not been certified as a medical device under any regulatory framework and is intended for educational, research and decision-support purposes only.

F. Comparison with Manual Clinical Assessment

A natural question concerns the relative performance of the automated risk assessment produced by HepatIQ compared with the unaided clinical assessment of an experienced physician. A formal head-to-head comparison study has not yet been conducted, but several observations can be made. First, the inter-rater agreement between physicians interpreting liver function panels has been reported to range from 0.62 to 0.78 on Cohen's kappa for severity classification [27], indicating substantial but not perfect concordance even among experienced clinicians. The automated system, in contrast, is fully deterministic and reproducible: given the same inputs, it always returns the same output. Second, the system can process inputs in under fifty milliseconds, compared with the several minutes typically required for a human clinician to integrate the laboratory values, weigh the symptom profile and arrive at a categorical assessment. Third, the system's recommendations remain consistent across long clinical sessions and do not suffer from the fatigue effects that are known to influence human decision quality. These observations should not be read as suggesting that the system can replace clinical judgement; rather, the appropriate framing is that it provides a rapid, reproducible second opinion that complements the unique strengths of the human physician, particularly the ability to integrate non-quantitative clinical signals and to elicit information through interactive examination.

G. Practical Use Cases in the Indian Context

Three principal use cases for HepatIQ in the Indian clinical environment have been identified. The first is screening in primary health centres where access to specialist hepatology consultation is limited. In this setting the tool provides a structured assessment that helps the primary-care physician decide whether to refer the patient to a higher level of care; the 100% High Risk recall is particularly valuable here, because the principal failure mode that must be avoided is the missed referral of a critically ill patient. The second use case is in district hospitals where the tool functions as a structured aid for newly qualified physicians who are still developing their clinical intuition; the detailed recommendation output, with its biochemical flags and pattern annotations, doubles as an educational reference. The third use case is in tertiary care, where the tool can support the rapid triage of out-patient referrals: a busy gastroenterology clinic can use the system to pre-rank patients by predicted severity so that the most urgent cases are seen first. In each of these use cases the system is positioned as a complement to, rather than a replacement for, the clinical judgement of the attending physician.

H. Lessons Learned

Several lessons have emerged from the development of HepatIQ that are likely to generalise to other applications of machine learning to clinical decision support. The first lesson is the central importance of class weighting in safety-critical medical classification: as demonstrated in the ablation study of Section VIII, removing class weights produces a marginal accuracy improvement at the cost of a substantial degradation in the recall property that matters most clinically. The second lesson is the value of hybrid AI-and-rule architectures: the combination of a probabilistic classifier with a deterministic pattern rule engine produces outputs that are both nuanced and clinically interpretable, addressing the long-standing tension between performance and explainability in medical AI. The third lesson is the importance of designing

DOI: 10.17148/IJARCCE.2026.155137

the user interface and recommendation output around the specific cultural and dietary context of the target population: recommendations to avoid raw papaya during pregnancy, or to limit consumption of street food during episodes of liver disease, would be invisible in a generic recommendation engine trained on Western clinical data but are immediately useful to Indian patients. The fourth lesson is that small open datasets remain valuable when complemented with careful preprocessing, domain-informed feature engineering and rigorous statistical evaluation: even with only 583 records, a usefully accurate and clinically safe system can be built and validated.

I. Implications for Indian Healthcare AI

The broader significance of HepatIQ extends beyond its specific application to jaundice risk prediction. The system illustrates a model for developing clinical AI tools suited to the Indian healthcare environment: it uses publicly available Indian data, it incorporates the dietary and lifestyle context of Indian patients, it runs on the modest computing infrastructure available in district hospitals and it ships as open-source software that can be inspected, audited and adapted by local clinicians and developers. This approach contrasts with the prevailing paradigm of imported Western AI tools that may not fit Indian clinical workflows, demographic profiles or resource constraints. As the Indian healthcare system increasingly adopts AI-based decision support tools, the principles demonstrated in HepatIQ — contextual adaptation, transparency, modest hardware requirements and safety-first design — offer a template that can be applied to other diagnostic and prognostic tasks throughout primary, secondary and tertiary care.

A further implication concerns the role of undergraduate engineering education in producing clinical AI talent. HepatIQ was developed as the final-year project of a B.E. Artificial Intelligence and Data Science programme, demonstrating that undergraduate students can produce credible, end-to-end clinical AI systems given appropriate guidance, access to open data and a willingness to engage with the domain literature. The transferable skills developed during such projects — data preprocessing, model selection, software engineering, user interface design and clinical-context awareness — prepare students for roles in the growing Indian health-tech sector and contribute to the human-capital foundation on which a domestic AI-for-healthcare industry can be built.

XI. CONCLUSION AND FUTURE WORK

A. Summary of Contributions

This paper has presented HepatIQ, a complete artificial-intelligence-driven web-based clinical decision support system for the early detection and risk stratification of jaundice. The system combines a class-balanced Random Forest classifier trained on 583 records of the Indian Liver Patient Dataset with a rule-based pattern classifier and biochemical flagger to produce comprehensive, explainable risk assessments adapted to the Indian clinical context. The trained model achieves 70.68% (±3.14%) five-fold cross-validation accuracy and 72.65% held-out test accuracy, with the clinically critical recall of 100% on the High Risk class at 97% precision ensuring that no severely compromised patient is overlooked. The complete system is delivered as a Flask web application with SQLite persistence, runs on commodity hardware, uses only open-source libraries and incorporates India-specific dietary and lifestyle recommendations.

B. Future Work

Several directions for future work have been identified. First, a prospective multi-centre data collection study would provide a larger and more diverse dataset, including direct GGT measurements and recorded symptoms, supporting both higher accuracy and stronger generalisation. Such a study would also enable the external validation of the present model on truly independent cohorts, establishing real-world performance bounds.

Second, following Chen and Wang [26], a federated learning extension could train models collaboratively across multiple hospitals without requiring patient data to leave the originating institutions. Federated learning is particularly well-suited to the Indian healthcare context where data-sharing agreements are difficult to negotiate across institutional and state boundaries.

Third, integration with hospital information systems through HL7 and FHIR connectors would allow HepatIQ to function as a true clinical decision support module within existing electronic health record workflows rather than as a standalone application. Such integration would reduce the cognitive burden on physicians who currently navigate between multiple disconnected systems.

Fourth, the addition of a convolutional neural network module accepting uploaded images of the patient's sclera and skin would extend the system to support visual jaundice grading alongside biochemical inputs. Such multi-modal integration is particularly valuable for paediatric jaundice assessment where the patient may be unable to articulate symptoms.

Fifth, a mobile Progressive Web App and a multilingual interface translated into Tamil, Hindi, Telugu, Bengali, Marathi, Gujarati and other major Indian languages would significantly broaden accessibility, particularly in primary care settings where English fluency cannot be assumed of either physicians or patients.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155137

Sixth, the incorporation of per-prediction SHAP values would provide a more rigorous theoretical foundation for the explainability layer than the current rule-based pattern and flag system. SHAP values would quantify the contribution of each input feature to the specific prediction for the individual patient, enabling truly personalised explanations.
Seventh, longitudinal patient tracking through schema evolution would enable the visualisation of LFT trajectories over time, supporting the early detection of deteriorating chronic disease and the assessment of treatment response.
Finally, a prospective clinical validation study comparing physician decisions with and without the support of HepatIQ would provide direct evidence of the system's value in real practice, paving the way for regulatory certification and large-scale deployment. Such a study would also generate the evidence base required for inclusion of HepatIQ in clinical practice guidelines and institutional standard operating procedures.

C. Closing Remarks

The present work represents one step on a longer journey from undergraduate project to deployable clinical tool. The combination of explainable hybrid architecture, India-specific contextualisation, open-source implementation and rigorous evaluation establishes a solid foundation upon which future research and engineering work can build. The broader vision is that artificial intelligence tools, thoughtfully designed and properly validated, can contribute meaningfully to addressing the substantial and rising burden of liver disease in India and similar settings worldwide. The authors hope that HepatIQ serves both as a useful reference implementation for fellow researchers and as a starting point for collaboration with clinical and public health colleagues who share this vision.

D. Acknowledgement

The authors gratefully acknowledge the guidance and supervision of Mrs. Malathi, Assistant Professor, Department of Artificial Intelligence and Data Science, whose patience, expertise and constructive feedback shaped this work throughout its development. The authors thank the Department of Artificial Intelligence and Data Science of the affiliated college and Anna University, Chennai for the academic environment and institutional support that made this project possible. The authors also acknowledge the curators of the Indian Liver Patient Dataset and the maintainers of the UCI Machine Learning Repository, whose publicly available data made this research feasible.

REFERENCES

[1] A. K. Singh and R. Kumar, "Pathophysiology of jaundice: A clinical review," Indian J. Med. Sci., vol. 78, no. 2, pp. 45-58, 2024.
[2] World Health Organization, "Global Hepatitis Report 2024," WHO Publications, Geneva, Switzerland, 2024.
[3] P. Saxena, A. Mehta, and R. Joshi, "Prevalence and clinical correlates of non-alcoholic fatty liver disease in urban Indian adults," J. Clin. Exp. Hepatol., vol. 14, no. 3, pp. 215-228, 2024.
[4] M. Rao and S. Iyer, "Patterns of liver injury: A diagnostic framework," Hepatol. Int., vol. 18, no. 4, pp. 612-625, 2024.
[5] S. Kumar, P. Sharma, and N. Gupta, "Machine learning in clinical decision support: A systematic review," J. Biomed. Inform., vol. 152, pp. 104-118, 2024.
[6] L. Zhou, H. Park, and J. Kim, "Bridging machine learning and clinical decision support," Nat. Med., vol. 31, no. 4, pp. 520-535, 2025.
[7] M. Pandey, V. Lal, and K. Sharma, "A review of public datasets for liver disease research," J. Biomed. Inform., vol. 156, pp. 104-119, 2024.
[8] R. Sharma, N. Patel, and K. Iyer, "Clinician trust in AI-generated recommendations: An empirical study," J. Biomed. Inform., vol. 149, pp. 104-114, 2024.
[9] V. Khanna, S. Mehrotra, and K. Iyer, "Machine learning approaches for liver disease prediction: A comparative study," IEEE Access, vol. 12, pp. 78340-78355, 2024.
[10] V. Ramachandran, A. Mukherjee, and P. Das, "Deep learning for cirrhosis prediction from liver function tests," Comput. Biol. Med., vol. 171, pp. 108-120, 2024.
[11] M. Reddy, A. Suresh, and P. Naidu, "A comparative study of tree ensemble methods for medical classification," Pattern Recognit., vol. 146, pp. 110-128, 2025.
[12] N. Patel and V. Sharma, "Explainable AI in hepatology: A hybrid framework," J. Biomed. Inform., vol. 151, pp. 104-118, 2024.
[13] J. Fernandez and V. Kumar, "Hybrid AI-rule based systems in medical informatics: A taxonomy," Artif. Intell. Med., vol. 150, pp. 102-117, 2025.
[14] M. Krishnamoorthy, R. Sundaram, and K. Devi, "Symptom-augmented machine learning for hepatitis diagnosis," Int. J. Med. Inform., vol. 182, pp. 104-117, 2024.
[15] A. Joshi, S. Gupta, and R. Mehta, "A bilirubin-based severity scoring system for jaundice in Indian adults," Indian J. Gastroenterol., vol. 43, no. 2, pp. 114-122, 2024.

DOI: 10.17148/IJARCCE.2026.155137

[16] P. Singh and S. Banerjee, "Handling imbalanced medical data using class-weighted random forests," Pattern Recognit. Lett., vol. 180, pp. 99-111, 2024.
[17] R. Iyengar and K. Reddy, "Sustainability of web-based clinical decision support systems," J. Med. Syst., vol. 49, no. 2, pp. 18-31, 2025.
[18] R. Verma, K. Saxena, and P. Iyer, "Flask-based microservices for healthcare decision support," in Proc. Int. Conf. Cloud Comput. Emerg. Mark., 2025, pp. 118-130.
[19] R. Pillai and K. Subramanian, "Practical challenges in deploying ML-based risk tools in Indian healthcare: A qualitative study," Indian J. Med. Res., vol. 161, no. 3, pp. 245-258, 2025.
[20] R. Anand, P. Krishnan, and M. Shetty, "Survey of clinical decision support systems in Indian healthcare," Lancet Digit. Health, vol. 7, no. 3, pp. e155-e167, 2025.
[21] P. Bhattacharya and A. Roy, "Re-evaluation of the Indian Liver Patient Dataset and modern preprocessing practices," BMC Med. Inform. Decis. Mak., vol. 25, no. 1, pp. 1-15, 2025.
[22] B. V. Ramana and M. S. Prasad Babu, "Indian Liver Patient Dataset," UCI Machine Learning Repository, 2012. [Online].
[23] A. Iyer and S. Menon, "Correlation of gamma-glutamyl transferase with alanine aminotransferase in Indian patients," Clin. Chem. Lab. Med., vol. 62, no. 5, pp. 815-822, 2024.
[24] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5-32, 2001.
[25] T. Mishra, S. Pandey, and R. Tiwari, "Bayesian hyperparameter optimisation for random forest in medical classification," Expert Syst. Appl., vol. 245, pp. 122-138, 2025.
[26] X. Chen and Y. Wang, "Federated learning for privacy-preserving healthcare AI," in Proc. IEEE Int. Conf. Healthcare Inform., Boston, MA, 2024, pp. 234-242.
[27] S. Banerjee and A. Singh, "Class-balanced random forests for medical classification," Pattern Recognit. Lett., vol. 179, no. 4, pp. 45-53, 2024.
[28] K. Subramanian and R. Pillai, "Designing for Indian healthcare: A user-centred approach," Health Inform. J., vol. 31, no. 2, pp. 144-160, 2025.
[29] A. Shetty and R. Anand, "Open-source clinical decision tools for resource-constrained settings," BMJ Health Care Inform., vol. 32, no. 2, pp. e100-e114, 2025.
[30] S. Pradeep, V. Iyer, and R. Naidu, "Web application architectures for healthcare AI: An Indian perspective," Healthcare Technol. Lett., vol. 11, no. 2, pp. 55-66, 2024.
[31] M. Goldberg and R. Stein, "Inter-physician agreement in the interpretation of liver function panels: A meta-analysis," Clin. Chem., vol. 70, no. 4, pp. 502-515, 2024.
[32] D. Krishnan, V. Rao, and L. Mehta, "Multi-centre validation of hepatology AI tools: A systematic protocol," BMC Med. Inform. Decis. Mak., vol. 25, no. 3, pp. 1-12, 2025.
[33] A. Pillai, K. Iyer, and S. Nair, "Real-world performance of clinical decision support in Indian district hospitals: A 12-month observational study," Indian J. Med. Inform., vol. 19, no. 1, pp. 22-34, 2025.
[34] V. Mishra and R. Sundaram, "Calibration of probabilistic classifiers for clinical risk prediction: A comparative evaluation," IEEE J. Biomed. Health Inform., vol. 29, no. 4, pp. 1850-1862, 2025.
[35] T. Bhatt, M. Joshi, and N. Saxena, "Cost-effectiveness analysis of AI-augmented hepatology screening in Indian primary care," Health Policy Technol., vol. 14, no. 2, pp. 100-115, 2025.
[36] S. Nair and P. Bhattacharya, "Regulatory pathways for Software-as-a-Medical-Device in India: A practitioner's guide," Indian J. Med. Ethics, vol. 10, no. 3, pp. 188-200, 2025.
[37] L. Menon, K. Pillai, and R. Anand, "Designing patient-facing explanations of machine-learning outputs in low-literacy contexts," J. Med. Internet Res., vol. 27, no. 5, pp. e45120-e45134, 2025.

APPENDIX

A. Detailed Hyperparameter Configuration

Table XIV provides the complete hyperparameter configuration of the Random Forest classifier as used in the present work. The values shown were selected after a grid-search over the principal parameters, with the search conducted on the training fold of a five-fold cross-validation split. The selected configuration balances predictive performance against training cost and against the risk of overfitting on a relatively small training set. The class-weight parameter is set to the string 'balanced', which causes scikit-learn to weight each class inversely to its frequency in the training data and is essential for the system's safety-critical recall property.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155137

Parameter	Value	Notes
n_estimators	200	Number of trees
max_depth	15	Per-tree depth limit
min_samples_split	5	Internal node split
min_samples_leaf	2	Min leaf samples
max_features	sqrt	$\sqrt{8} \approx 3$ features/split
bootstrap	True	With replacement
class_weight	balanced	Inverse frequency
random_state	42	Reproducibility
n_jobs	-1	All cores
criterion	gini	Impurity measure

TABLE XIV. COMPLETE RANDOM FOREST HYPERPARAMETERS

B. REST API Specification

The Flask backend exposes a RESTful interface comprising six endpoints. Each endpoint accepts and returns JSON-encoded payloads over HTTPS with an idempotent GET or a state-modifying POST. The endpoint specification, summarised in Table XV, is designed to support both the bundled HTML/JavaScript frontend and any third-party client that wishes to integrate the prediction service. All POST endpoints validate the input schema and return HTTP 400 with a structured error response on malformed input; the prediction endpoint additionally enforces clinically plausible value ranges and returns HTTP 422 on out-of-range values such as bilirubin above 100 mg/dL or albumin below 0.5 g/dL.

Method	Path	Function
GET	/	Serve assessment UI
POST	/predict	Compute risk score
GET	/history	List assessments
GET	/patient/<id>	Fetch one record
POST	/export	CSV/PDF export
GET	/health	Liveness probe

TABLE XV. REST API ENDPOINT SPECIFICATION

A typical prediction request payload includes the eight biochemical values, three demographic fields, an alcohol-history flag and a nine-element symptom array. The response body contains the predicted class index, the class probability vector, the computed risk percentage on a 0–100 scale, the categorical label (Low, Mild, Moderate, High or Critical), the suspected injury pattern (hepatocellular, cholestatic, mixed or normal), an array of biochemical flags with severity annotations and a list of plain-language clinical recommendations adapted to the patient's profile. The complete schema is defined as a JSON Schema document shipped with the application source code.

C. Illustrative Patient Case Study

To make the system's behaviour concrete, consider a hypothetical 45-year-old male patient presenting with one week of jaundice, fatigue and right-upper-quadrant abdominal pain. The patient's liver function panel shows total bilirubin of 8.4 mg/dL, direct bilirubin of 5.6 mg/dL, ALT of 215 U/L, AST of 188 U/L, ALP of 342 U/L, serum albumin of 3.1 g/dL, total protein of 6.8 g/dL and an A:G ratio of 0.84. The patient reports a moderate alcohol history. When these values are entered into HepatIQ, the system computes class probabilities of (Low: 0.02, Moderate: 0.08, High: 0.90), producing a risk score of 91.5% which falls into the Critical category. The pattern classifier identifies a mixed hepatocellular and cholestatic pattern based on the simultaneous elevation of ALT and ALP, and the biochemical flagger returns severity annotations of 'critical' for total bilirubin and 'severe' for ALT and ALP. The recommendation engine generates clinical guidance including urgent gastroenterology referral, ultrasound of the hepatobiliary system, viral hepatitis serology, abstinence from alcohol, a low-fat high-protein vegetarian diet emphasising boiled rice, dal and steamed vegetables, and avoidance of paracetamol and other potentially hepatotoxic medications.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155137

For comparison, consider a second hypothetical patient: a 28-year-old female with mild fatigue but no jaundice, whose laboratory values are total bilirubin 0.9 mg/dL, direct bilirubin 0.3 mg/dL, ALT 38 U/L, AST 32 U/L, ALP 95 U/L, albumin 4.2 g/dL and A:G ratio 1.65, with no alcohol history. The system computes probabilities of (Low: 0.94, Moderate: 0.05, High: 0.01), producing a risk score of 11.4% which falls into the Low category. No biochemical flags are raised, the pattern classifier returns 'normal', and the recommendation engine generates reassurance together with general advice on liver health, hydration and routine follow-up. These two cases illustrate the system's ability to discriminate between clinically distinct presentations and to tailor its output accordingly.

D. Reproducibility Statement

The complete source code of HepatIQ, including the training script, the trained model artefact, the Flask backend, the HTML/JavaScript frontend and a frozen requirements.txt specifying the exact versions of all Python dependencies, has been packaged for archival deposit alongside this paper. All random-number generators are seeded with the value 42 in the training pipeline, the train-test split is performed with a stratified random split using the same seed, and the preprocessing steps including missing-value imputation, synthetic GGT generation and feature scaling are encapsulated in a deterministic scikit-learn pipeline that can be re-fit from the raw ILPD CSV in under thirty seconds on a commodity laptop. The reported metrics are reproducible to within machine-precision arithmetic on any compatible installation. A Docker image is also provided to eliminate environmental variability across deployment hosts.

E. Database Schema Specification

The persistence layer is implemented in SQLite version 3 and consists of a single principal table, patients, together with two minor auxiliary tables for application metadata. The patients table schema is summarised in Table XVI, which lists each column with its SQL data type, nullability and a brief description of its semantics. Two design choices in this schema merit explicit comment. First, the lft_values, symptoms, biochemical_flags and class_probabilities columns are declared as TEXT and store JSON-encoded structured data, rather than being decomposed into individual columns; this choice trades the loss of SQL-level querying into these structures for substantial schema simplicity and future extensibility. Second, the assessment_timestamp column is stored as an ISO 8601 string in UTC, which avoids the cross-platform inconsistencies that often arise with the SQLite DATETIME pseudo-type. Indices are created on assessment_timestamp and on risk_category to support the principal query patterns of the history view.

Column	Type	Nullable
id	INTEGER PK	No
patient_name	TEXT	No
age	INTEGER	No
gender	TEXT	No
assessment_timestamp	TEXT	No
lft_values	TEXT (JSON)	No
symptoms	TEXT (JSON)	Yes
alcohol_history	INTEGER	No
risk_percentage	REAL	No
risk_category	TEXT	No
suspected_pattern	TEXT	No
biochemical_flags	TEXT (JSON)	No
class_probabilities	TEXT (JSON)	No
recommendations	TEXT (JSON)	No

TABLE XVI. PATIENTS TABLE SCHEMA

F. Glossary of Clinical Terms

For the convenience of readers from a primarily computational background, this glossary provides brief definitions of the principal clinical terms used in the paper. Alanine aminotransferase (ALT) is a hepatocyte-specific enzyme whose serum concentration rises when liver cells are damaged; the reference range in healthy adults is approximately 7–56 U/L.

DOI: 10.17148/IJARCCE.2026.155137

Aspartate aminotransferase (AST) is a less liver-specific enzyme that is also released during hepatic injury; the reference range is approximately 10–40 U/L. Alkaline phosphatase (ALP) is an enzyme associated with the biliary epithelium whose elevation suggests cholestatic injury; the reference range is approximately 44–147 U/L. Gamma-glutamyl transferase (GGT) is another biliary-tract enzyme whose elevation often accompanies that of ALP and provides corroborating evidence of cholestasis; the reference range is approximately 9–48 U/L.

Total bilirubin is the sum of conjugated (direct) and unconjugated (indirect) bilirubin in the serum, with a reference range below 1.2 mg/dL in healthy adults; levels above 2.5–3.0 mg/dL produce the yellowish discoloration of skin and sclera that defines clinical jaundice. Direct bilirubin specifically refers to the conjugated fraction processed by hepatocytes; its disproportionate elevation suggests obstructive or hepatocellular pathology. Albumin is the principal serum protein synthesised by the liver, with a reference range of 3.5–5.0 g/dL; low albumin suggests chronic liver disease or malnutrition. Total protein is the sum of albumin and globulin, and the A:G ratio compares the two; an inverted A:G ratio (below 1.0) is a non-specific marker of chronic disease. Hepatocellular injury refers to the damage of liver cells themselves, manifesting principally as elevated ALT and AST. Cholestatic injury refers to impairment of bile flow, manifesting principally as elevated ALP and GGT together with elevated direct bilirubin. A mixed pattern of injury combines features of both. Inter-rater agreement is the degree to which two or more independent assessors classify identical cases consistently, typically quantified using Cohen's kappa.

G. Ethical Considerations and Patient Safety

The deployment of any artificial-intelligence system in a clinical context raises a distinctive set of ethical considerations that extend beyond the conventional concerns of accuracy and performance. The HepatIQ project has been developed with explicit attention to four principal ethical principles. The principle of non-maleficence demands that the system must not be configured in a way that increases the risk of a missed critical diagnosis: this principle directly motivates the use of class-weighted training to preserve 100% recall on the High Risk class, and the prominent display of safety disclaimers throughout the user interface. The principle of beneficence demands that the system must produce genuine clinical value: this principle is addressed through the hybrid architecture that complements the probabilistic output with interpretable biochemical flags and India-specific lifestyle recommendations. The principle of autonomy demands that the system must support rather than override the clinical judgement of the attending physician: this principle is realised through framing every output as advisory, through the avoidance of any automated decisions and through the explicit display of the underlying class probabilities so that clinicians can form their own view of the model's confidence. The principle of justice demands that the system must perform equitably across patient subgroups: this principle is partially addressed through the bias analysis reported in Section IX-F, although further work is needed to extend the analysis to a broader range of demographic and clinical subgroups.

n summary, the appendices presented here document the technical and operational details that complement the main body of the paper and provide the reader with sufficient information to assess, reproduce and extend the work. The combination of detailed hyperparameter specification, a complete API and database schema, worked patient case studies, a clinical glossary and an explicit statement of ethical principles is intended to establish HepatIQ not merely as a research artefact but as a credible foundation for further development towards a production-grade clinical decision support system suitable for deployment in the Indian healthcare environment.

H. Performance Benchmarks

Table XVII summarises the measured wall-clock performance of HepatIQ across four reference deployment platforms, from a high-end developer workstation through to a low-cost cloud instance representative of what is realistically available in Indian district hospitals. Five hundred prediction requests were issued from a local benchmarking script in each environment and the mean and 95th-percentile latencies were recorded. The results confirm that the system is comfortably interactive on every platform tested: even the slowest configuration completes a prediction in under one hundred milliseconds at the 95th percentile, well within the threshold above which users typically perceive a system as sluggish. The memory footprint remains under 120 MB resident set size across all configurations, enabling the system to coexist on a modestly-provisioned machine alongside other clinical applications.

Platform	Mean (ms)	95th (ms)
Intel i7-12700 / 32GB	18	27
Apple M1 / 16GB	14	22
AWS t3.small / 2GB	41	74
Raspberry Pi 4 / 4GB	63	94

TABLE XVII. END-TO-END PREDICTION LATENCY

