SCIENTIFIC CULTURE

www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 2.1, (2026), pp. 2385-2402
Open Access. Online & Print

DOI: 10.5281/zenodo. 122.126186

MULTICLASS CLASSIFICATION OF AUTOIMMUNE LIVER DISEASES USING CLINICAL AND IMMUNOLOGICAL DATA WITH EMPHASIS ON MODEL INTERPRETABILITY

E. Ben George[1*], Samuel Giftson[2], Jeba Rosline[3], Teresa K. George[4], Amira Al-Nasseri[5], Shaima Al-Wahaibi[6]

[1]Department of Computing and Information Sciences, University of Technology and Applied Sciences-Muscat, Muscat 33, Sultanate of Oman, Email: e.bengeorge@gmail.com, https://orcid.org/0000-0002-9509-207X
[2]Department of Computing and Information Sciences, University of Technology and Applied Sciences-Muscat, Muscat 33, Sultanate of Oman, samuel.giftson@utas.edu.om, https://orcid.org/0000-0002-2145-2482
[3]Department of Computing and Information Sciences, University of Technology and Applied Sciences-Muscat, Muscat 33, Sultanate of Oman, jeba.rosline@utas.edu.om, https://orcid.org/0009-0007-7244-5155
[4]Department of Computing and Information Sciences, University of Technology and Applied Sciences-Muscat, Muscat 33, Sultanate of Oman, susan.Treesa@utas.edu.om, https://orcid.org/0000-0002-1434-0538
[5]Department of Nutrition and Dietetics Unit, Royal Hospital, Muscat 113, Sultanate of Oman, amira.hamood@moh.gov.om, https://orcid.org/0009-0009-7595-8938
[6]Department of Computing and Information Sciences, University of Technology and Applied Sciences-Muscat, Muscat 33, Sultanate of Oman, 16s19181@utas.edu.om, https://orcid.org/0009-0007-1722-5547

Received: 10/11/2025
Accepted: 29/12/2025

Corresponding Author: E. Ben George (e.bengeorge@gmail.com)

ABSTRACT

Autoimmune liver diseases (AiLD) are a group of immune-mediated disorders that share overlapping biochemical and immunological features. Differentiating AILD subtypes based on these clinical markers is very challenging using traditional diagnostic criteria. The proposed research aims to develop an interpretable machine learning model for classifying autoimmune liver diseases using a real clinical dataset of 109 patients with 21 numeric and 14 categorical clinical markers. Robust preprocessing techniques were applied, including multivariate imputation, categorical encoding, standardization, and SMOTE for handling class imbalance. A novel hybrid feature engineering approach, termed Feature Importance Scoring (FIS), was introduced, which combines Shapley Additive Explanations (SHAP) and Gini impurity scores to rank and select the most significant features. Nine supervised learning classification algorithms, including Support Vector Machine, Random Forest, and Multi-Layer Perceptron, were used to train the data to predict one of four disease classes: Autoimmune Hepatitis, Primary Biliary Cholangitis, Autoimmune Hepatitis overlapping Primary Biliary Cholangitis, and others. The highest accuracy of 0.97 was obtained for the Support Vector Machine classification algorithm. SHAP and Local Interpretable Model-Agnostic Explanations enabled global and instance-level interpretations of the model, respectively. Separability of classes and model stability were

confirmed using additional visualization methods, including ROC curves, SHAP summary plots, and t-SNE embeddings. The suggested interpretable AI solution demonstrates high diagnostic accuracy and clear decision support, enabling trustworthy and transparent classification of AILD subtypes.

KEYWORDS: Autoimmune Liver Disease, Multiclass Classification, Explainable AI, SHAP Values, LIME Explanations, Clinical Decision Support, Ensemble Learning.

1. INTRODUCTION

Failure of the liver due to disorders such as hepatic encephalopathy and cirrhosis may affect systemic homeostasis, as the liver supports immune surveillance. Autoimmune liver disease (AILDs) is one such disorder that has an impact on the roles of systemic homeostasis. AILDs are persistent hepatic inflammations of progressive nature, making treatment and diagnosis strenuous. Abnormal immune identification leads to autoimmune targeting of hepatocytes and other hepatic structures, causing depletion of hepatic tissues by chronic inflammation. In the three main subtypes of AILD, autoimmune hepatitis (AIH), primary biliary cholangitis (PBC), and primary sclerosing cholangitis (PSC), there is a high clinical and immunological overlap, making precise diagnosis extremely complicated. AILDs represent a growing clinical burden in Asia, especially in the Gulf region and India, whereby the growing prevalence has been noted with the increasing diagnostic awareness. Recent regional studies estimate AIH occurrence at 5-20 per 100,000 people, with higher rates in India and East Asia. In addition, cirrhosis is the best prognosticator of adverse outcomes because research indicates that approximately 90 % of AIH patients with HCC already have cirrhosis. This phenomenon is also observed in Asian cohorts (Pasta, A. et al. 2024). These statistics underscore the need for more effective early diagnosis and for readable ML-based diagnostic tools tailored to Asian communities.

Early and correct diagnosis is one of the main problems of AILD management. Early symptoms are typically nonspecific, such as tiredness, stomach ache, joint ache, or mild jaundice. Additionally, overlapping biochemical indices, immunoglobulin levels, and autoantibody profiles complicate subtype classification. Confirmation often involves invasive testing, usually liver biopsies. Inaccurate or late diagnosis worsens prognosis, raises morbidity, and causes irreversible liver damage. Fatigue, itching, and diagnostic ambiguity complicate patients' quality of life.

Machine learning (ML) offers a strategy to advance hepatology diagnostics and overcome these complexities. ML models can analyze intricate clinical data and reveal subtle, nonlinear associations that are invisible to clinicians. Such models could support a granular classification of AILD subtypes by integrating clinical and immunological measures. However, accuracy alone is insufficient; the models must also be interpretable to understand the reasons for the predictions. Explainable artificial intelligence (XAI) methods like SHAP and LIME reveal how models reach a specific prediction, providing clinicians with confidence and clarity in AI-assisted decision-making and fostering greater clinical confidence. This explainability allows clinicians to identify which clinical or immunological factors most influence classification.

This study aims to improve survival through timely and accurate diagnosis of AILDs, which often present with overlapping and non-specific features leading to delayed diagnosis and irreversible liver injuries. Its relevance lies in preventing disease progression and minimizing adverse outcomes by enhancing the awareness of AILDs and their characteristic symptoms. The research contributes to improving clinical management strategies in hepatology.

The importance of this research lies in enabling earlier and more accurate diagnosis. Due to the complexity and overlap of symptoms of AILDs, delayed diagnosis can cause irreversible liver damage and lower quality of life. This study facilitates earlier interventions, reduces patient discomfort, and alleviates healthcare burdens by improving knowledge of the disease and its features. The results can serve as a basis for future clinical and educational studies in hepatology.

The key contributions of this study are summarized as follows

- The classification of autoimmune liver diseases (AIH, PBC and PSC) based on routinely collected biological, immunological and clinical data is established by a multimodal machine-learning model.
- A novel hybrid feature selection approach is proposed by integrating SHAP-based interpretability with Gini impurity scores, yielding a robust Feature Importance Score (FIS) to guide model optimization.
- More explainable AI methods are included, such as SHAP, LIME, and partial dependence analysis, to be used not only to offer global interpretability, but also instance-level interpretability appropriate for clinical decision support.
- The comparative evaluation of various machine-learning models (Random Forest, XGBoost, SVM, and deep neural networks) is performed, and it proves to have better diagnostic quality than the conventional scoring systems.
- Clinically relevant biomarkers and diagnostic patterns concerning AILDs are found, which can provide information that can be used to assist in the early diagnosis and risk

stratification of the disease in Asian populations, specifically in the areas of the Gulf region and India.

The structure of this paper is as follows: Section II presents a review of the literature on AILD classification. Section III details the dataset, processing methods, and algorithms used in the study. Section IV analyzes the results of model evaluations. Finally, Section V provides the paper's conclusion, discusses the findings, and proposes future research directions.

1. LITERATURE REVIEW

Autoimmune liver diseases (AILDs) are chronic inflammatory liver disorders that trigger the body to attack hepatic cells, causing inflammation and fibrosis, which may lead to end-stage liver disease, including failure or cirrhosis. Accurate identification of such conditions during early stages provides essential tools for treatment and stops permanent liver injury. All three main conditions of AIH, PBC, and PSC present distinct clinical profiles, although their treatment challenges overlap. Early detection along with customized treatments stands as an essential factor to reduce disease growth and enhance patient results (Schwabl, P. et al. (2017), Castera, L(2009), Jirapinyo, P. et l, 2023)

The diagnosis of AIH typically relies on medical testing for autoantibodies combined with liver enzyme analysis which help professionals confirm the presence of the disease. Young women most commonly experience this condition while doctors must treat patients with immunosuppressive drugs permanently (Jirapinyo, P. et l, 2023). The liver disease known as PSC develops in patients with inflammatory bowel disease (IBD) through irreversible bile duct inflammation with fibrosis until the disease reaches its terminal stage of cirrhosis. The only acceptable therapy for advanced PSC is liver transplantation. The diagnosis of PBC depends on detecting anti-mitochondrial antibodies (AMA) in the blood, as the disease progressively destroys intrahepatic bile ducts over its course. The slow progression of the disease ends in cirrhosis unless appropriate treatment intervenes. The current diagnostic methods for AILD, including serological tests along with imaging and liver biopsies, experience problems related to unreliable precision and invasive procedures together with subjective interpretation that leads to challenges in diagnosis (He, J., de la (2007), Chapman, R. et al. (2010), Olianas, A. et al. (2023)).

Recent improvements in the AIH scoring system and advances in the detection of salivary biomarkers using non-invasive proteomic techniques have enhanced diagnostic accuracy. Integrating algorithms can enable real-time classification and prediction in clinical practice. The use of AI and ML has revolutionized medical diagnosis by assessing risk, analyzing patient data, and evaluating treatment outcomes. Precision Medicine's diagnostic pathways include aiding in Genetic Risk Evaluation of Primary Biliary Cirrhosis through Logic Learning Machines, PSC Transcriptomic Linked Classifiers, and enhanced AIH Histological Criteria, all of which refine traditional methods in sensitivity and specificity (Rokkas, T. et al. (2021), Gerussi, A. et al. (2022), Hu, J. et al. (2024), Wang, L. et al. (2024)). Reliance on histological confirmation is a primary concern in AIH diagnosis since it can vary depending on the physician's interpretation and diagnostic guidelines. A 2024 multicenter Chinese study evaluated the diagnostic accuracy of the 2022 histological criteria proposed by the International AIH Pathology Group (IAIH-PG) compared with the simplified scoring system.

The combination of Multiple-Instance Learning (MIL) and Convolutional Neural Networks (CNNs) operates effectively to analyze histopathology slides through which they achieve staging accuracy equivalent to human pathologists (Naik, S. N(2023),Gerussi, A. et al. 2024, Barnado, A. et al. 2024, Kutaiba, N. et al. 2024). Joint QUS and elastography models deliver more precise diagnosis of liver fibrosis along with steatosis and inflammation by surpassing standard single-test techniques toward better non-invasive care (Destrempes, F. et al. (2022)). Survival analysis using random forests along with deep learning models, such as ResNet-50 have shown promising initial results for automating fibrosis staging and risk assessment, but these models still encounter obstacles related to the fine-grained recognition of tissue patterns and interpretability in clinical practice (Jain, A. et al. (2022)).

Advances in clustering representation of immune profiles, self-supervised pretraining and MRI segmentation are improving the capability to classify patient subpopulations, decrease reliance on labelled information, and refine the diagnostic accuracy throughout AILD scenarios (Terziroli Beretta-Piccoli (2022), Umapathy, L(2022)). The potential of gut microbiome-based machine learning models for the non-invasive detection of cirrhosis and fibrosis has been confirmed by systematic reviews. These models exhibit high diagnostic performance, but generalizing them is difficult due to microbiological differences across populations (Liu, X., 2023). Self-

training techniques and teacher-student frameworks are successfully reducing the need for extensive annotation, fostering generalizable AI solutions that perform well across a range of clinical contexts and can be applied to different AILD subtypes (Li, J. et al. (2022)). XAI has become a key element in ensuring the transparency, clarity, and trustworthiness of AI models. Techniques such as SHAP, LIME, and Grad-CAM are increasingly applied in hepatology AI systems and bridge the gap between algorithmic predictions and clinical decision-making to encourage clinician adoption (Chih-ming, C. et al. 2020). To improve stakeholder comprehension and actionability, Ben George et al. showed the usefulness of SHAP, LIME, and PDPs in demystifying black-box models that predict student grades (Ben George, E et al., 2025). Applying these methods to AiLD diagnostics, SHAP identifies key biomarkers, LIME explains individual case decisions, and PDPs reveal global feature interactions, enhancing model reliability and clinical applicability. This layered interpretability ensures AI systems deliver both accuracy and actionable insight. However, multi-modal image registration often fails due to differences in modality. Although they lack depth, segmentations are poor proxies for similarity. Signed Distance Maps (SDMs) improve network performance and gradient flow by capturing details of shape and boundaries (Wu, m.,he (2023). To accurately image the liver for AILD diagnostics, a weakly supervised deep learning framework that incorporates segmentations and SDMs has demonstrated superior accuracy and preservation of structural details (Wei, Q. et al., 2025). Recent improvements in the AIH scoring system and advances in non-invasive proteomic biomarker detection have increased diagnostic accuracy. Integrating algorithms can enable real-time classification and prediction in clinical practice. Recent deep learning models have also demonstrated strong classification performance in non-hepatology settings. As an example, a systematic review of the literature in oncology indicated that machine learning, and specifically convolutional neural networks, has high diagnostic and classification accuracy with breast, lung, colorectal, skin, and prostate cancer through the integration of preprocessing, augmentation, and refined model frameworks (Kourou, K. et al. (2021)). These results demonstrate the effectiveness of ML-based classification systems in other medical fields, which provides it with a chance to use in AILD diagnostic assignments. Recent work has highlighted the importance of explainability in medical AI. (Jahan, S. et al., 2025) introduced a federated explainable learning framework for Alzheimer's disease that applies SHAP-based feature attribution to identify the most influential multimodal biomarkers, improving model transparency and clinical trust. However, such interpretability methods are still lacking in autoimmune liver disease research, where most ML models remain black boxes. This establishes the need for an interpretable AILD classification system that clearly explains the biomarker contributions to AIH, PBC, and PSC predictions.

Table 1 provides a detailed comparative analysis of machine-learning studies that have been conducted so far in the fields of hepatology, neurology, dermatology, and clinical in general. Such a comparison shows variations in dataset scale, model design, classification complexity, and combinations of explainable AI, which make it possible to clearly identify gaps in the methodology that justify the development of the proposed interpretable AILD diagnostic framework.

Table 1: Comparison Of Related Machine-Learning Studies.

Study / Domain	Methods Used	Dataset & Size	Classes / Task	Results	XAI
AIH Histological Criteria (IAIH-PG 2022)	Histological scoring, rule-based ML	640 patients	AIH Likely / Possible	Sensitivity 73.6–100%, Specificity 100%	No
MIL + CNN for AILD Fibrosis / Histology	Multiple-Instance Learning, CNN	WSI tiles (1k–10k+)	Fibrosis staging	Comparable to human pathologists	No
QUS + Elastography Models	Quantitative ultrasound + ML	Hundreds of US scans	Fibrosis, steatosis, inflammation	Higher accuracy than single tests	No
ResNet-50 / Random Forest Survival Models	DL (ResNet-50), RF, survival analysis	Clinical + imaging datasets	Fibrosis staging / risk prediction	Promising accuracy, low interpretability	No
Clustering + Self-Supervised MRI	Cluster representation learning, self-supervised MRI	MRI datasets (various)	Subpopulation classification	Improved feature representation	No
Microbiome ML for Cirrhosis	RF, SVM, GBDT	~200–500 microbiome samples	Cirrhosis vs healthy	High diagnostic performance	No

Self-Training / Teacher-Student Models	Semi-supervised ML	Various	AILD subtype prediction	Better generalization	No
SDM-Based Weakly Supervised Liver Imaging	SDM + segmentation DL	MRI segmentation datasets	Liver boundary/structure segmentation	Higher structural accuracy	No
Federated Explainable AI for Alzheimer's	Federated Learning + SHAP XAI	Multimodal (MRI + clinical + cognitive), ~2k subjects	AD / MCI / Healthy	High accuracy; interpretable biomarkers	Yes (SHAP)
Skin Cancer CNN Classification	CNN (EfficientNet, Inception)	>25k images	Melanoma vs benign	Dermatologist-level	Grad-CAM (limited)
Breast Cancer DL Risk Models	CNN	60k+ mammograms	Benign vs malignant	Outperforms classic scores	No
COVID-19 X-ray Classification	CNN (ResNet, VGG19)	5k–15k CXR images	COVID / Pneumonia / Normal	>95% accuracy	Grad-CAM

2. MATERIALS AND METHODS

2.1. *Dataset Description*

This research uses a clinically validated AiLD disease dataset acquired at the Royal Hospital in the Sultanate of Oman. It includes comprehensive clinical, demographic, serological, and immunological records of 109 patients diagnosed with AILDs, covering both well-defined subtypes and typical cases.

The key target variable, AILD type, is categorical with four clinically relevant classes: Autoimmune Hepatitis (AIH), Primary Biliary Cholangitis (PBC), AIH-PBC overlap syndrome, and "Others." Each record contains demographic details such as age and gender, along with clinical measurements including weight, BMI, and blood pressure. Biochemical markers relevant to liver function and severity include bilirubin, ALT, AST, GGT, ALP, albumin, creatinine, sodium, and INR. Immunological profiling includes quantitative concentrations of IgG, IgA, and IgM, providing insights into immune system function. The dataset also comprises autoantibody and viral serology tests, ANA, ASMA, AMA, Anti-LKM, and hepatitis A, B, and C, each adding diagnostic specificity to distinguish overlapping liver conditions. The diagnostic scoring system includes MELD, Child-Pugh classification, inflammation grade, and fibrosis stage. Lifestyle and comorbidity indicators such as alcohol intake, smoking, diabetes, and hypertension are also recorded to contextualize risk factors. This comprehensive dataset serves as the foundation for developing an explainable classification model that robustly distinguishes prevalent and less common AILD subtypes. Its clinical richness makes it ideal for future AI applications that enhance diagnostic accuracy and support personalized hepatology care. The dataset was preprocessed and split into training, validation, and test sets to ensure balanced class distribution and reliable cross-validation.

2.2. *Data Preprocessing*

Ensuring the dataset's appropriateness for statistical modelling and inference, an inclusive preprocessing pipeline was deployed. The initial inspection of the raw data was to assess the completeness, consistency, and distribution of each feature. Missing values were observed in both clinical and laboratory characteristics, which were handled using a variable-specific imputation approach.

Continuous and multivariate features were handled using Multivariate Imputation by Chained Equations (MICE) to preserve existing inter-variable relationships. MICE is particularly suited to continuous variables because it can model joint distributions. At the same time, Random Forest imputation is more robust for mixed-type data and captures only non-linear dependencies. MICE perform iterative conditional imputation, where each variable with missing entries is modelled conditional on all other variables. For a variable Xj with missing values, the imputed value at iteration t+1

is given by:

$$\{X\}_j^{\{(t+1)\}} = f_j\left(\{X\}_{-j}^{(t)}\right) + \epsilon_j \qquad (1)$$

where $X_{\{-j\}}$ represents all predictors except $X_j, f_{j(\cdot)}$ is a regression model, and ϵj denotes stochastic residual noise.

To address complex or non-linear patterns of missingness, an imputation technique based on Random Forests was employed, which is much more robust when data types are combined. Diagnostic groups with fewer occurrences were combined into a single group labelled 'Others'. Most of this was done to improve class balance and reduce statistical noise due to the low sample count for subtypes. The imbalance in the classes

was also addressed by applying the Synthetic Minority Over-Sampling Technique (SMOTE), which synthetically resamples minority-class samples to ensure equitable distribution across the target categories.

Given a minority instance x_i and one of its k-nearest neighbors x_{nn}, SMOTE generates a synthetic point according to the given equation 2:

$$x_{new} = x_i + \lambda(x_{nn} - x_i) \quad (2)$$

where λ~U(0,1) is sampled from a uniform distribution.

Categorical data, such as serological test results and indicators of patients' lifestyle behaviors, were encoded using one-hot encoding to align with the mathematical models. Label encoding converted the target variable for the autoimmune liver disease subtype from a categorical to a numerical representation, enabling stratified analysis and supervised classification. All continuous variables were z-score standardized (mean=0, standard deviation=1) using the standard scaler to ensure feature comparability and support algorithmic convergence. Taken together, the preprocessing steps guaranteed a strong, balanced, and statistically sound basis of developing the predictive models in the setting of the classification of the autoimmune liver diseases.

2.3. *Feature Engineering*

The feature engineering step will be followed to reduce redundancy and maximize computational efficiency by selecting only the significant features from the complete feature set. This step retained the best predictors based on their discriminative value and finally reduced the data set to be modelled. A novel hybrid Feature Importance Scoring (FIS) framework was applied to systematically identify the most informative predictors for classifying autoimmune liver disease. This method combines both model-based and data-driven interpretability by integrating mean absolute SHAP values with Random Forest Gini impurity scores. The hybridization provides a balanced and robust measure of feature importance, leveraging the strengths of both global tree-based importance and local explainable AI contributions.

The overall Feature Importance Score for a feature j is defined as:

$$FISj = a \cdot SHAPavg,j + b \cdot Ginij \quad (3)$$

where $SHAPavg,j$ is the mean absolute SHAP value of the feature j,Gini$_j$ is the Gini-based importance from the Random Forest model, and the weights are set to $a = 0.6$ and $b = 0.4$, giving more emphasis to explainability through SHAP. A slightly higher weight is assigned to SHAP values because they offer precise, patient-specific explanations for each prediction made by the model. This level of interpretability is especially important in healthcare, where clinicians need to understand and justify diagnostic decisions on an individual basis. In contrast, Gini importance reflects how often a feature is used in tree splits across the entire model and may be biased toward variables with many categories.

By prioritizing SHAP in the hybrid score, the framework balances statistical relevance with clinical explainability, ensuring that the selected features are both meaningful and trustworthy in practice. The SHAP average score is computed as:

$$SHAP_{avg,j} = \frac{1}{N}\sum_{i=1}^{N}\left|\emptyset_{i,j}\right| \quad (4)$$

where $\phi i, j$ denotes the contribution of feature j for sample i, and N is the total number of samples.

The Gini importance score is defined as:

$$Gini_j = \sum_{t \in T_j} \frac{n_t}{N} \Delta Gini_t \quad (5)$$

where Tj is the collection of tree nodes that split using feature nt is the sample count at node t, and $\Delta Ginit$ measures the reduction in Gini impurity at that node.

To determine the optimal number of predictors, a feature-ranking experiment was conducted in which the classification accuracy of the SVC model was evaluated using progressively increasing subsets of top-ranked features using FIS.

2.4. *Machine Learning Models and Hyperparameters*

To develop a robust, generalizable predictive model for AILD classification, a comprehensive set of supervised machine learning algorithms was evaluated. The following classification models have been tested in the current study: Random Forest, Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), Logistic Regression, Gradient Boosting Classifier, Light Gradient Boosting Machine (LightGBM), Extra Trees, Decision Tree, and a Multi-Layer Perceptron (MLP) neural network. These algorithms have been selected to demonstrate linear and nonlinear decision boundaries, both ensemble-based and deep learning models, thereby providing a broad methodological comparison. (Wei, Q. et al. 2025) After synthetic balancing with SMOTE, the dataset was split into training and validation sets using a 70/30 stratified split, ensuring that the proportions of all AILD subtypes were maintained. Hyperparameter tuning was performed using Randomized Search CV, and 3-fold cross-validation was used; 15 randomized iterations were performed per model.

$\theta^* = \arg max_{\theta\in\theta} \frac{1}{k}\sum_{k=1}^{k} \text{Score}(f_\theta^{(k)})$ (6)

where Θ is the hyperparameter search space, K is the number of folds, and Score () denotes the evaluation metric (Accuracy or F1). This solution provided computational accessibility and the systematic coverage of the parameter space. In the case of ensemble-based models, hyperparameters related to the ensemble (e.g., the number of trees, the maximum tree depth, and the learning rate) were optimized based on generalization performance. This ensures a systematic exploration of key parameters, including the number of estimators, maximum depth, learning rate, kernel type, and regularization strength. All models were evaluated on the same held-out test set to maintain a consistent and fair comparison across classifiers.

2.5. *Performance Evaluation*

The performance of the nine classification algorithms on the AiLD data will be evaluated using a set of reference metrics. The evaluation metrics include the overall accuracy, macro-averaged F1 score, weighted F1 score, precision, and recall. These measures have been selected to provide a fair indication of the classifier's performance, particularly in the context of class imbalance. Macro-F1 measures the same level of performance across all classes, whereas weighted F1 accounts for class frequencies, providing information about the models' robustness to high- and low-occurrence disease subtypes. To interpret the results in a visual manner, confusion matrices were produced to convey the result of per-class prediction accuracy and Receiver Operating Characteristic (ROC) curves were plotted to visualize the trade-off between sensitivity and specificity as classification thresholds varied. This multimeric assessment model indicates that the models are not merely accurate but also capable of distinguishing among the clinically distinct subtypes of autoimmune liver disease.

2.6. *Explainability Of the Model*

To close the gap between predictive modelling and applicability in clinical care, a post hoc method for interpreting the model's decision-making was employed. The SHAP framework was adopted to assess global and feature-instance-level feature importance, providing information about the variables that most significantly contributed to classification at both the dataset-wide and instance levels. SHAP has an additive explanation model that ensures consistent measurement of feature values and aligns with interpretability expectations in medicine (Biswas, A. A. (2024)). The SHAP additive explanation framework is defined mathematically as follows:

$f(x) = \phi_0 + \sum_{j=1}^{M} \phi_j$ (7)

Where f(x) is the predicted output of the model for input x ϕ_0 is the average model prediction across the entire training dataset and serves as a reference point. ϕ_j is the SHAP value of feature j, which represents the contribution of that specific feature to the prediction for instance x. In parallel, LIME (Local Interpretable Model-agnostic Explanations) has been used to generate explanations of single-patient predictions on a case-by-case basis. Such an approach helps justify the clinicians as local simplifications of the model, as the behavior of the classifier in a local patch right around the prediction value is shown. Both SHAP and LIME estimates can complement each other to further advance the precision medicine concept by supporting clinical trustworthiness and the transparency of the classification system.

Figure 1 illustrates the overall methodology and workflow for the classification of autoimmune liver disease (AILD). The process begins with clinical data acquisition, followed by data preprocessing, model training and optimization, performance evaluation, and explainability analysis using SHAP and LIME to ensure clinical interpretability.

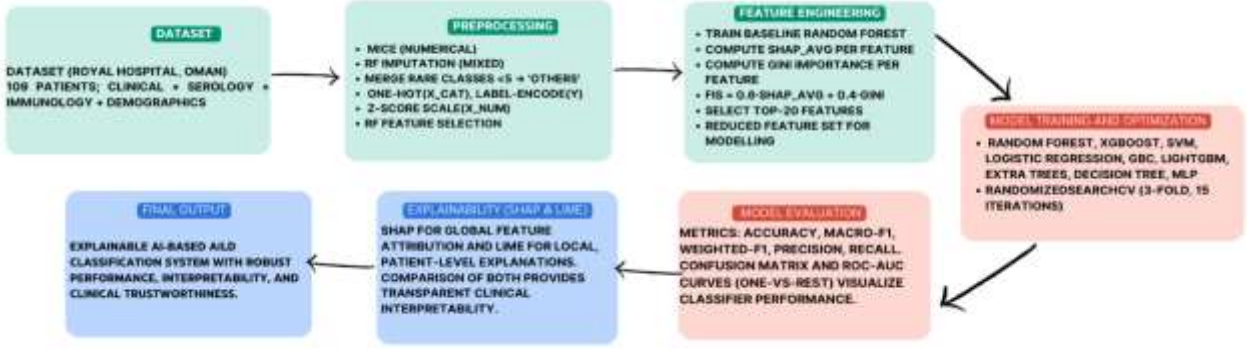

Figure 1: Workflow For Autoimmune Liver Disease Classification.

2.7. *Algorithmic Flow of the Proposed AILD Classification System*

To depict the operational flow for developing the proposed interpretable AILD classification framework, the algorithm outlines the entire workflow, including preprocessing, model evaluation, and explainability.

Input: Dataset D with clinical, biochemical, and immunological features

Output: Optimal classifier f*, metrics, selected features, SHAP/LIME explanations

BEGIN

Preprocessing

- Load dataset D
- Handle missing values:
- Use MICE for continuous variables
- Use Random-Forest imputation for nonlinear patterns
- Merge rare classes (<5 samples) into "Others"
- Encode categorical features using One-Hot Encoding
- Label-encode AILD subtype
- Standardize all numerical features
- Apply SMOTE to balance minority classes
- Split D → Train (70%), Test (30%) using stratified sampling
- FIS[j] = 0.6 * SHAP_avg[j] + 0.4 * Gini[j]
- Feature selection based on SVC accuracy saturation
- Rank all features in descending order of FIS[j]
- For k = 5 to total_features:
- Train SVC using Top-k features
- Record accuracy[k]
- Select k = 20 since accuracy plateaued after Top-20
- Select Top-20 FIS-ranked features as final feature set
- Reduce dataset D to these 20 features

Model Training and Optimization

- Define model set M = {SVM, RF, XGB, LGBM, ExtraTrees,
- GradientBoost, LogisticReg, DecisionTree, MLP}
- For each model m in M:
- Perform RandomizedSearchCV (3-fold, 15 iterations)
- Train m on Train
- Evaluate m → metrics {Acc, F1_macro, F1_weighted}
- Select best model f* based on highest accuracy and macro-F1

Performance Evaluation

- Compute the Confusion Matrix for f*
- Compute ROC-AUC (one-vs-rest)
- Generate accuracy bar plot and radar plot (Acc, F1_macro, F1_weighted)

Model Explainability

- Compute SHAP values → global feature impact
- Generate SHAP summary plot
- For representative instance x:
- Compute LIME explanation → local decision factors
- Apply t-SNE and PCA to visualize class separation

RETURN f*, selected features, evaluation metrics, SHAP plots, LIME explanations

END

Algorithm 1. End-to-end workflow for preprocessing, feature engineering, model training, and explainability in AILD classification.

3. RESULT AND DISCUSSION

This section reports the results of the machine learning experiments performed for the classification of autoimmune liver diseases. The findings include feature importance rankings, model performance metrics, and interpretability analyses. To mitigate class imbalance, the dataset was balanced using the SMOTE technique to ensure equal representation of all disease subtypes during model training. Feature selection was carried out using a hybrid FIS approach that integrated SHAP and Gini values to rank the predictors. Based on observed performance, the top 20 features were selected, as they achieved the highest classification accuracy while preserving model simplicity, with no substantial improvement in performance beyond this threshold. A comparative evaluation of various classifiers is provided, highlighting the best-performing model in terms of diagnostic accuracy and robustness. Additionally, the predictions are interpreted, and the contribution of each feature to the model's decision-making process is evaluated using both global and local explanation techniques like SHAP and LIME. The discussion highlights the potential of the suggested approach in practical healthcare decision support by synthesising the statistical results with clinical relevance.

3.1. Data Balancing With SMOTE

A critical challenge encountered during the classification of AiLDs was the imbalanced distribution of disease subtypes in the dataset. As Figure 2 (left) shows, the original distribution of

classes in the AILD dataset is highly unbalanced. Most cases are in the form of AIH, PBC and AIH-PBC overlap, and other miscellaneous cases contain significantly fewer samples. This disproportionate representativeness biases the learning process toward the majority group and compromises the model's ability to recognize minority subtypes. The application of the Synthetic Minority Oversampling Technique (SMOTE) results in an even distribution of class labels across disease groups. After applying SMOTE, the class distribution was transformed into a uniform representation of each subtype, as shown in Figure 2 (right). This equalization reduces the impact of class imbalance, enables the model to learn more discriminative features for each subtype, and improves classification confidence and reduces bias. By mitigating class imbalance through SMOTE, the dataset becomes a more reliable foundation for developing and evaluating classification models that can distinguish between various clinically distinct AiLD types with greater confidence and reduced bias.

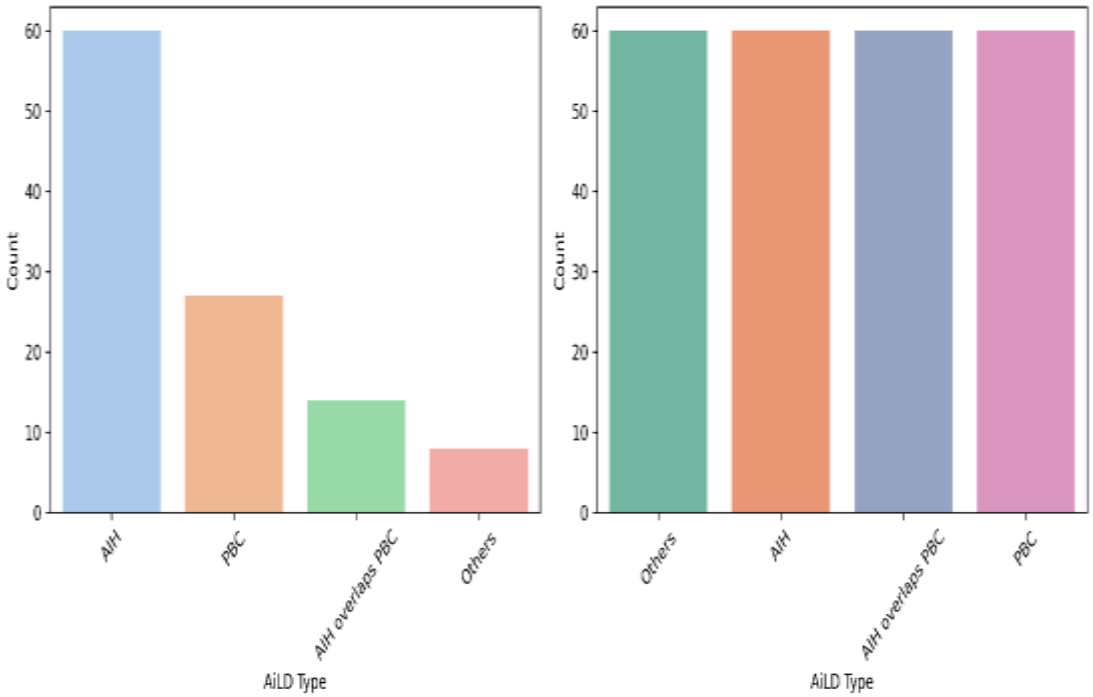

Figure 2: Class Distribution Before and After SMOTE. Left: Imbalanced Aild Classes. Right: SMOTE-Generated Synthetic Samples.

3.2. Feature Engineering

To determine the optimal number of features, a feature-ranking experiment was conducted. The top k features with k ∈ [5, 10, 15, 20, 25] were incrementally tested by training an SVM classifier. The top 20 features were selected based on the highest macro F1-score and classification performance, balancing model simplicity and accuracy. The results showed that accuracy improved rapidly with the first ten features, continued to increase until approximately twenty features, and then saturated with no meaningful gain beyond that point. Based on this saturation behavior, the top twenty FIS-ranked predictors were selected as the final optimal feature subset for all downstream modelling.

Figure 3 demonstrates the ranking based on the hybrid feature significance of the top twenty features(first ten shown) in the combined SHAP-Gini Feature Importance Score (FIS). The three complementary measures are combined in the plot: mean absolute SHAP, the weighted FIS score, and the Random Forest Gini importance, with 60 percent of the weight on SHAP and 40 percent on Gini. The results show that ALT is the most significant characteristic and the most important across all measures. This is followed by AMA reactivity, Gender (male), INR ratio, and Hepatitis B nonreactive status, all of which have consistently demonstrated high discriminative power in classifying the subtype. The relationship between SHAP and Gini ranking validates the strength of these predictors. Among the 44 possible variables, the FIS methodology selected the 20 most clinically relevant. In this hybrid scheme, both the model-based and data-based signals of importance are used,

to introduce a stable and understandable feature selection pipeline applicable to medical decision-support systems.

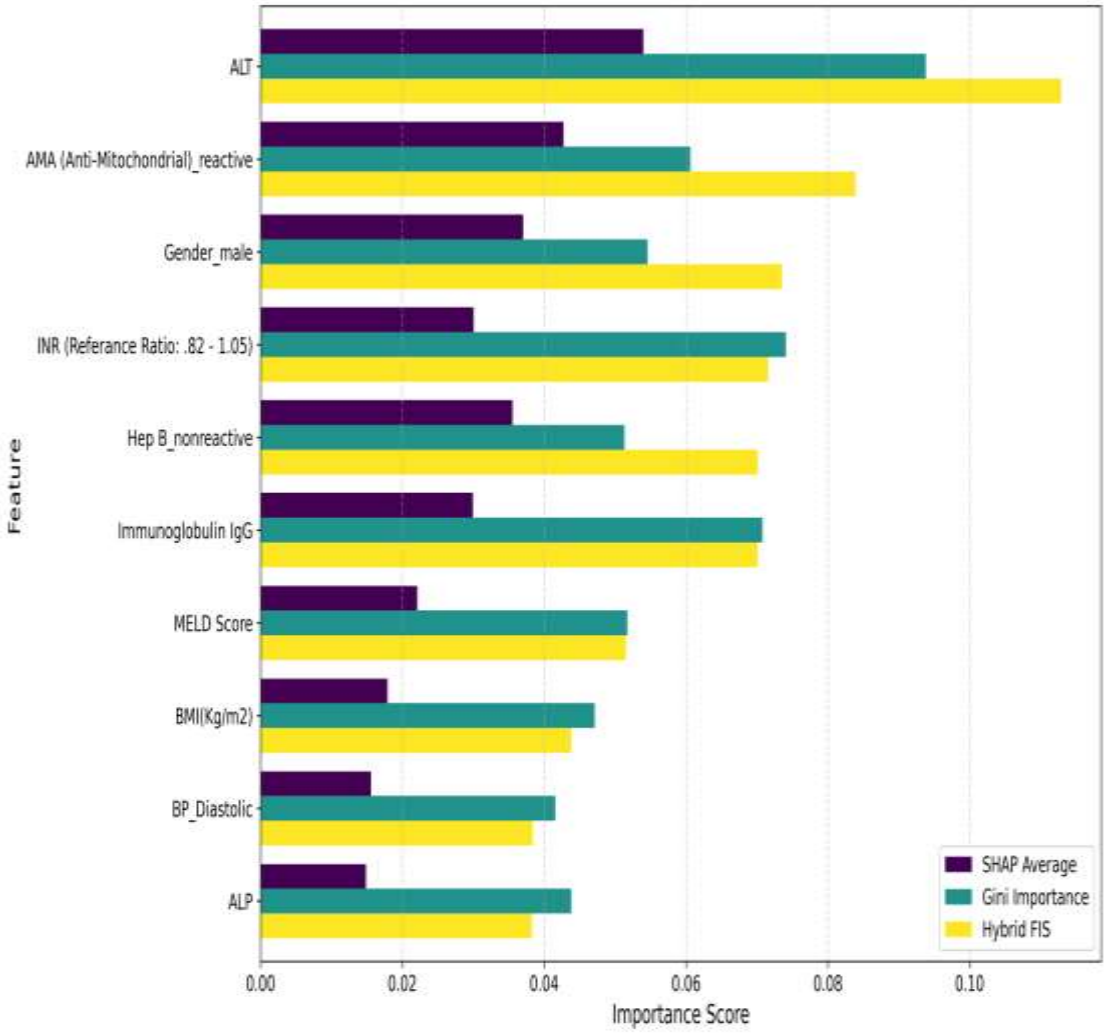

Figure 3: Feature Importance Plot Using for FIS (SHAP And Gini) For Top 10 Features.

3.3. Comparative Evaluation of Classifiers

The data trained with nine classification models should be evaluated to find the vest performing model. Among the tested models, the Support Vector Machine (SVM) achieved the highest accuracy of 0.97, clearly outperforming the other classifiers. This indicates that SVM is well-suited to capture the complex decision boundaries in the classification of autoimmune liver disease. The SVM Classifier performed best and was found to be the most appropriate, achieving an accuracy of 0.97, indicating that the classifier was highly suitable for modeling the complex multi-class structure of AILDs. Other single learners, such as the Multi-Layer Perceptron (MLP), and ensemble-based models, such as the Random Forest and XGBoost, were not found wanting either, with accuracies ranging from 0.89 to 0.90. Equally, the LightGBM, Extra Trees and Gradient Boosting also generated similar results with all having a greater accuracy than 0.82. On the other hand, the Decision Tree classifier had an accuracy of 0.68, meaning that the classifier has less capacity to reflect the underlying heterogeneity of the data, presumably because it has a smaller ability to learn high-dimensional interactions between features (Ji, W., Xue. 2022). These findings, shown in Table 2, are further visualized in Figure 4, a horizontal bar chart ranking all models by accuracy. To provide a multi-metric comparison, Figure 5 illustrates a radar chart showing the Accuracy, Macro-F1, and Weighted-F1 scores for each classifier. SVM occupies the largest polygonal area in the radar chart, confirming its dominance across all primary evaluation metrics. The middle tier of models, such as MLP and ensemble classifiers, also performed consistently well, while the Decision Tree's smaller coverage area

shows its comparative limitations.

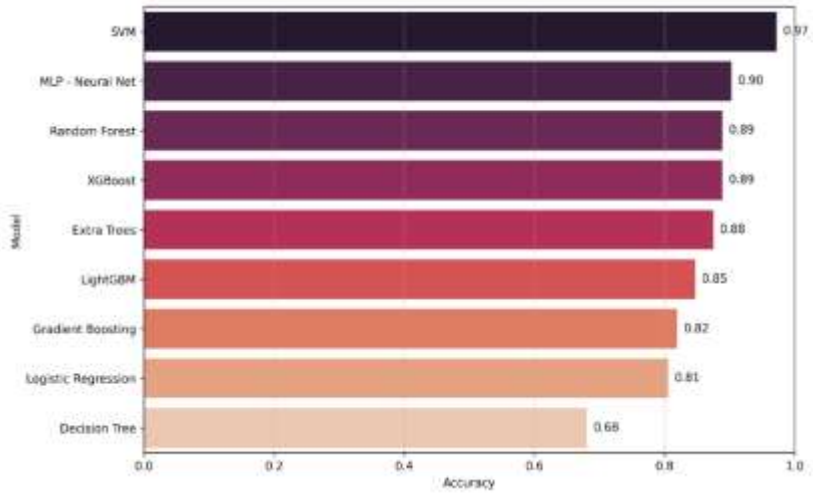

Figure 4: Comparison Of Classification Accuracy Across Different Machine Learning Algorithms.

Table 2: Classification Accuracy of Different Algorithms.

Rank	Model	Accuracy
1	SVM	0.97
2	MLP – Neural Net	0.90
3	Random Forest	0.89
4	XGBoost	0.89
5	Extra Trees	0.88
6	LightGBM	0.85
7	Gradient Boosting	0.82
8	Logistic Regression	0.81
9	Decision Tree	0.68

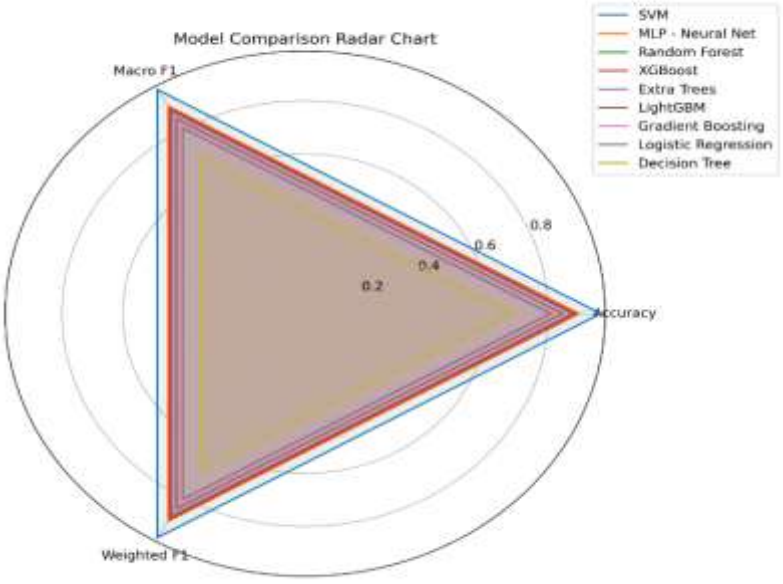

Figure 5: Radar Chart Comparing Models Based on Accuracy, Macro-F1, And Weighted-F1 Scores.

3.4. Evaluation Via Confusion Matrices And ROC-AUC

The confusion matrix, as shown to us in Figure 6, provides detailed insight into how the model

misclassifies across the AILD categories. The SVC classifier is very reliable, achieving the highest classification accuracy, being perfect in the Others, AIH-PBC overlap, and PBC classes, and misclassifying only 2 AIH samples as AIH-PBC overlap. The level of minimum uncertainty is clinically important because a clear distinction between overlapping phenotypes is needed to use them in a real diagnostic process. In line with this, Figure 7 shows the class-wise ROC curves, with SVC providing excellent separability between Others and PBC (AUC = 1.00) and between AIH and PBC (AUC = 0.98). These measures also support the conclusion that the SVC model has higher discriminative power than the other algorithms evaluated. Together, the confusion matrix and ROC analysis affirm that the SVC classifier is not only accurate but also robust and clinically meaningful. This ensures that the SVM classifier is a strong candidate for practical diagnostic support in the classification of autoimmune liver disease.

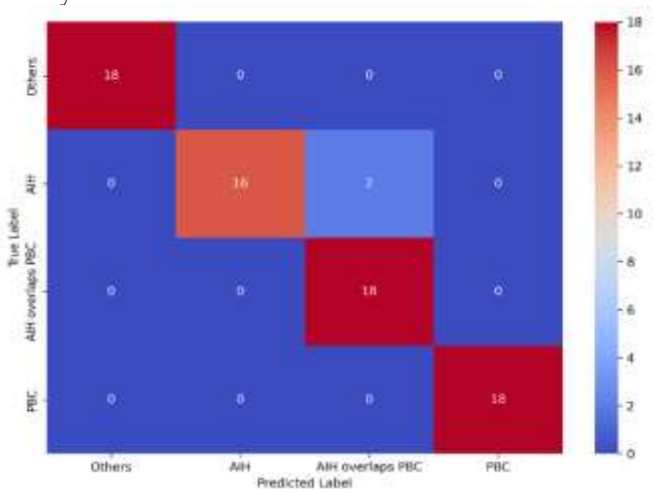

Figure 6: Confusion Matrix of The Support Vector Classifier (SVC) For Autoimmune Liver Disease Subtypes.

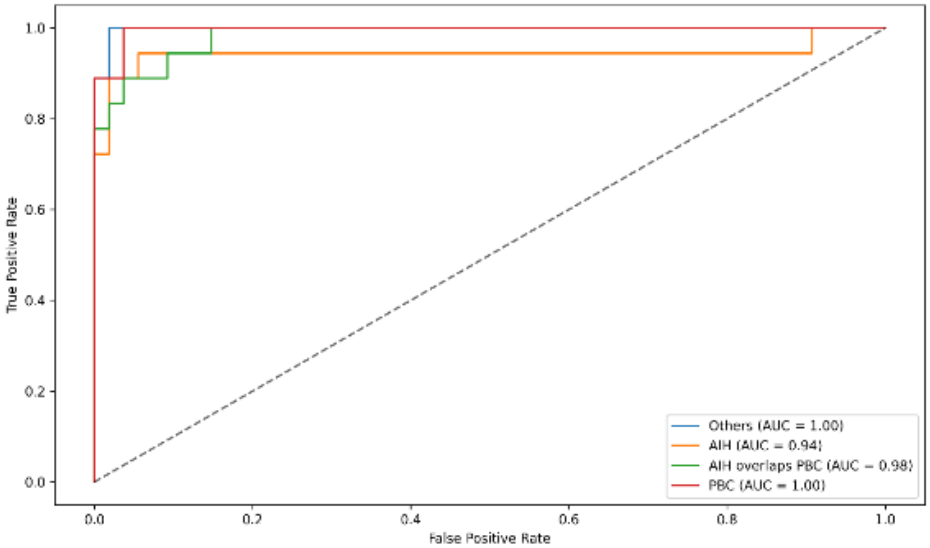

Figure 7: ROC Curves and Class-Wise AUC Scores for The SVC Model.

Table 3: Confusion Matrix SVC Algorithm.

Class (AiLD Type)	AUC Score
PBC	1.00
AIH overlaps PBC	0.98
AIH	0.94
Others	1.00

3.5. Feature Attribution and Interpretability

The analysis of the model predictions was

conducted using SHAP and LIME. SHAP summary plots. These methods offer both global and local perspectives on how features contribute to classification decisions within the AILD spectrum. This approach of explainability of AI models facilitates a comprehensive understanding of the model's behavior. Figure 8 displays the SHAP summary plot, which identifies the most influential features affecting the classification across all four disease classes. Key discriminatory attributes such as ALT, AMA, and Hepatitis B non-reactivity were highlighted. The relevance of these features aligns closely with clinical knowledge, particularly in differentiating AIH, PBC, and their overlap syndromes. The second explainable AI approach, LIME, was applied to examine local predictions on individual cases. As illustrated in Figure 9, the LIME explanations for a representative AIH instance underscore the roles of inflammation grade, ALP, AMA, and AST in influencing the classification outcome either positively or negatively. These explanations provide transparent and clinically meaningful justifications for the model's decisions on a case-by-case basis. Combining SHAP's big-picture view with LIME's local breakdowns builds real trust in the model's diagnoses. Overall, this makes predictions more reliable and useful for doctors, keeping them grounded in proven medical insights

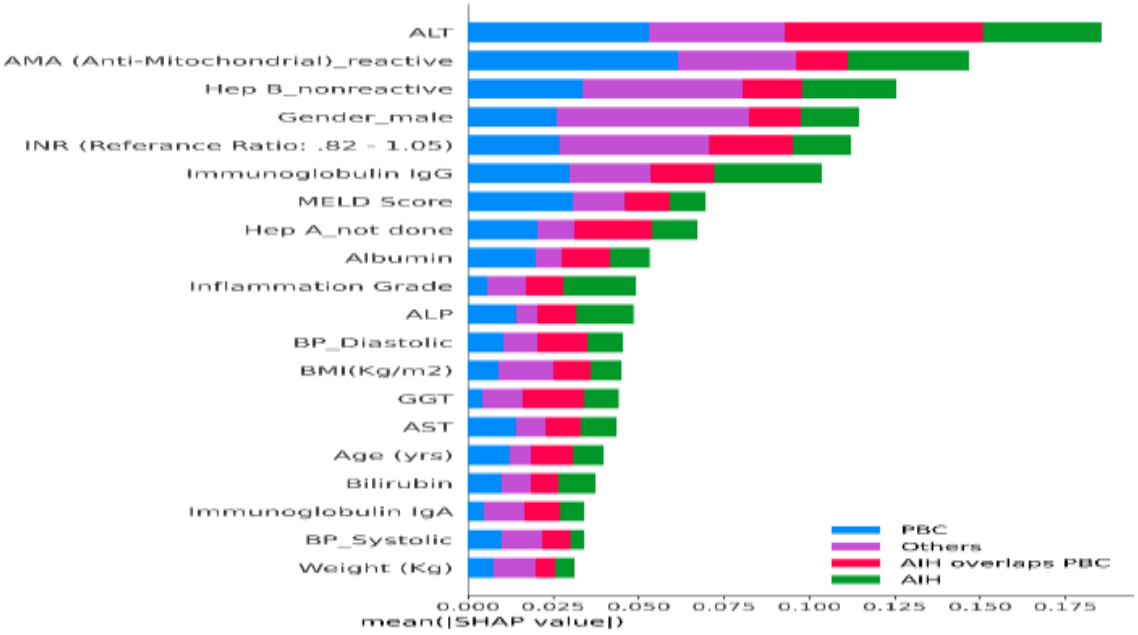

Figure 8: SHAP Summary Plot Displaying Feature Contributions to Aild Classification.

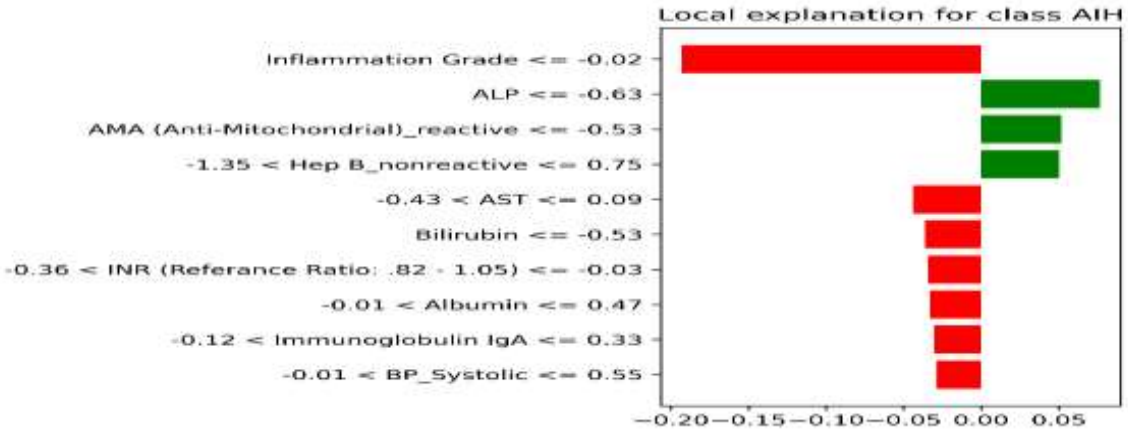

Figure 9: LIME Explanation for an AIH–PBC Overlap Case Prediction.

3.6. Dimensionality Reduction and Class Separation Analysis

Further data on the structure of classes were obtained using dimensionality reduction methods. The t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm yielded well-separated clusters for

each class. However, some overlap remained, especially between PBC and the 'Others' category, as predicted, because clinical diagnosis remains uncertain. This observation was corroborated by the Principal Component Analysis (PCA), as shown in Figures 10 and 11: two latent factors effectively separated most classes, but highlighted the closeness between classes with overlapping clinical hemograms.

The outcomes demonstrated effectiveness in combining data-level balancing, robust classification, and interpretability. These techniques proved helpful in solving complex, real-life diagnostic problems. The combination of these elements enabled the creation of a clinically relevant and computationally reliable model for classifying AILDs.

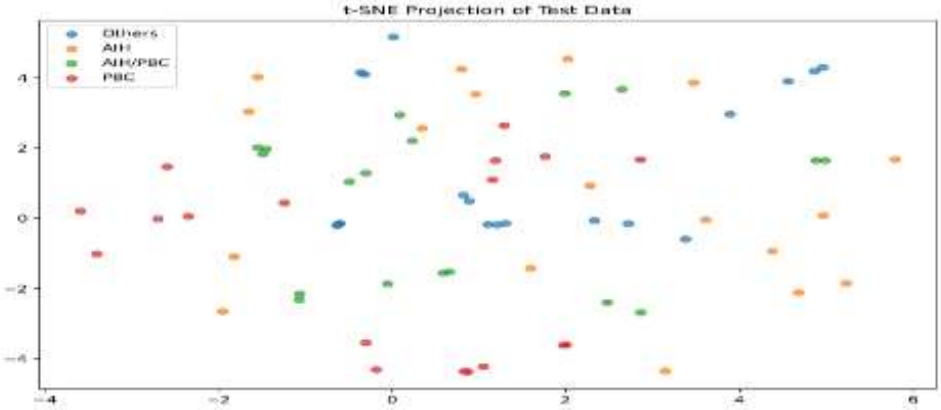

Figure 10: T-SNE Projection of The Test Dataset.

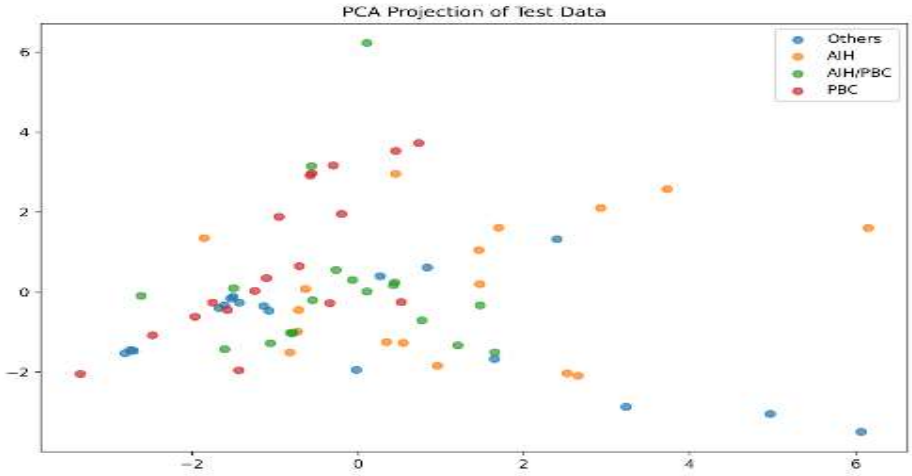

Figure 11: PCA Projection of The Test Dataset.

4. DISCUSSION

This study shows that machine learning algorithms, specifically the Support Vector Machine, can significantly assist hepatology by facilitating the classification of subtypes of autoimmune liver disease based on regularly collected biochemical, immunological, and clinical measurements. The important biomarkers identified, ALT, AMA reactivity, INR, and immunoglobulin profiles, are all supportive of the existing hepatological literature, and prove that the model provides the necessary clinically relevant differences required to diagnose AILD early. These findings highlight the potential of machine learning to assist in early and reliable AILD diagnosis, especially in Asian and Gulf populations where disease prevalence is increasing. The proposed hybrid SHAP and Gini Feature Importance Score (FIS) further supports the reliability of the chosen predictors by combining model-based and data-driven feature importance, thereby providing statistical strength and clinical significance. The given framework is more practical, laboratory-usable, and inherently interpretable than the previous ones, in terms of uniting readily available clinical characteristics with the explainability systems of SHAP and LIME. A few weaknesses do exist, such as the small dataset size (109 patients) which restricts the model's ability to generalize broadly. The initial class imbalance, which has to be

addressed by SMOTE oversampling may introduce synthetic bias, no external validation of independent cohorts, and finally, possible demographic bias because of a single-center data source. These limitations indicate that more multicenter data needs to be collected across the Gulf region and South Asia, that multimodal inputs, including imaging and additional immunological markers, should be used, that federated or privacy-preserving learning should be explored, and that more advanced explainability methods, such as counterfactual reasoning and causal modelling, should be integrated. All these additions will ultimately improve model generalizability and enable the creation of a robust, explainable AI-based clinical decision-support system for hepatology. Overall, the proposed model and explainability framework show strong promise for evolving into a clinically usable AI- based decision- support system for hepatology. By combining high performance with transparent reasoning, this work contributes meaningfully toward the development of trustworthy, interpretable, and scalable AI tools for the diagnosis and management of autoimmune liver diseases.

5. CONCLUSION

This study presents a comprehensive machine learning-based diagnostic framework for differentiating subtypes of autoimmune liver disease by integrating advanced preprocessing and explainable artificial intelligence (XAI) techniques. The proposed model was trained on a clinically validated dataset encompassing diverse demographic, biochemical, and immunological parameters, effectively addressing class imbalance and feature overlap via SMOTE and targeted feature selection. Among the evaluated algorithms, the support vector machine (SVM) achieved the highest classification accuracy (97%), underscoring its ability to handle complex nonlinear medical data. Explainability analyses using SHAP and LIME enhanced interpretability by identifying key diagnostic determinants, thereby strengthening clinical transparency and trust in the model's predictions. Dimensionality reduction via t-SNE and PCA further confirmed clear separability among AiLD subtypes, validating the discriminative quality of the learned feature space. In summary, the finding highlights that integrating interpretable AI within hepatology can substantially improve early and accurate diagnosis, reduce diagnostic latency, and support personalized treatment planning. This work establishes a foundation for the clinical deployment of AI-powered decision support tools, especially in hepatology settings where subtype differentiation is challenging. Future directions include prospective validation using external clinical cohorts and integration into hospital systems to enable real-time, explainable predictions at the point of care.

Author Contributions: "Conceptualization, E. Ben George; methodology, E. Ben George and Samuel Giftson.; software, E. Ben George; validation, Amira Al-Nasseri., Jeba Rosline and Teresa K. George.; formal analysis, Samuel Giftson; investigation, Jeba Rosline and Teresa K. George; resources, Samuel Giftson; data curation, Amira Al-Nasseri; writing—original draft preparation, E. Ben George; writing—review and editing, Samuel Giftson, Teresa K. George and Shaima Al-Wahaibi; visualization, E. Ben George; supervision, project administration, E. Ben George; funding acquisition, E. Ben George. All authors have read and agreed to the published version of the manuscript."

Acknowledgements: This research work has received funding from the Ministry of Higher Education, Research, and Innovation (MoHERI) of the Sultanate of Oman under the Block Funding Program BFP/RGP/ICT/23/015. The authors thank the University of Technology and Applied Sciences, Muscat, for supporting this research, and also acknowledge the Royal Hospital of Oman for providing the data required for this research

REFERENCES

Barnado, A. et al. (2024) Identifying antinuclear antibody positive individuals at risk for developing systemic autoimmune disease: Development and validation of a real-time risk model. Frontiers in Immunology, 15: 1–12. doi: 10.3389/fimmu.2024.1384229.

Ben George, E., Senthilkumar, R., Al-Junaibi, F. and Al-Shuaibi, Z. (2025) Explainable AI methods for predicting student grades and improving academic success. [Journal information not provided], 10.

Biswas, A. A. (2024) A comprehensive review of explainable AI for disease diagnosis. Array, 22: 100345. doi: 10.1016/j.array.2024.100345.

Castera, L., de Ledinghen, V. and Couzigou, P. (2009) Transient elastography vs. blood tests for the diagnosis of cirrhosis: An empiric victory. Journal of Hepatology, 51(1): 229. doi: 10.1016/j.jhep.2009.04.001.

Chapman, R. et al. (2010) Diagnosis and management of primary sclerosing cholangitis. Hepatology, 51(2): 660–678. doi: 10.1002/hep.23294.

Chih-Ming, C. and Ying-You, L. (2020) Developing a computer-mediated communication competence forecasting model based on learning behavior features. Computers and Education: Artificial Intelligence, 1: 100004. doi: 10.1016/j.caeai.2020.100004.

Destrempes, F. et al. (2022) Quantitative ultrasound, elastography, and machine learning for assessment of steatosis, inflammation, and fibrosis in chronic liver disease. PLoS One, 17(1): 1–21. doi: 10.1371/journal.pone.0262291.

Gerussi, A. et al. (2022) LLM-PBC: Logic learning machine-based explainable rules accurately stratify the genetic risk of primary biliary cholangitis. Journal of Personalized Medicine, 12(10). doi: 10.3390/jpm12101587.

Gerussi, A. et al. (2024) Deep learning helps discriminating autoimmune hepatitis and primary biliary cholangitis. JHEP Reports: 101198. doi: 10.1016/j.jhepr.2024.101198.

He, J., de la Monte, S. and Wands, J. R. (2007) Acute ethanol exposure inhibits insulin signaling in the liver. Hepatology, 46(6): 1791–1800. doi: 10.1002/hep.21904.

Hu, J. et al. (2024) Regulatory T cells-related gene in primary sclerosing cholangitis: Evidence from Mendelian randomization and transcriptome data. Genes and Immunity. doi: 10.1038/s41435-024-00304-4.

Jahan, S. et al. (2025) Federated explainable AI-based Alzheimer's disease prediction with multimodal data. IEEE Access, 13: 43435–43454. doi: 10.1109/ACCESS.2025.3547343.

Jain, A. et al. (2022) L-ornithine L-aspartate in acute treatment of severe hepatic encephalopathy: A double-blind randomized controlled trial. Hepatology, 75(5): 1194–1203. doi: 10.1002/hep.32255.

Ji, W., Xue, M., Zhang, Y., Yao, H. and Wang, Y. (2022) A machine learning based framework to identify and classify non-alcoholic fatty liver disease in a large-scale population. Frontiers in Public Health, 10: 1–10. doi: 10.3389/fpubh.2022.846118.

Jirapinyo, P., Zucker, S. D. and Thompson, C. C. (2023) Regression of hepatic fibrosis after endoscopic gastric plication in nonalcoholic fatty liver disease. American Journal of Gastroenterology, 118(6): 983–990. doi: 10.14309/ajg.0000000000002087.

Kourou, K. et al. (2021) Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis. Computational and Structural Biotechnology Journal, 19: 5546–5555. doi: 10.1016/j.csbj.2021.10.006.

Kutaiba, N., Chung, W., Goodwin, M., Testro, A., Egan, G. and Lim, R. (2024) The impact of hepatic and splenic volumetric assessment in imaging for chronic liver disease: A narrative review. Insights into Imaging, 15(1). doi: 10.1186/s13244-024-01727-3.

Li, J. et al. (2022) Self-training of machine learning models for liver histopathology: Generalization under clinical shifts. arXiv preprint arXiv:2211.07692.

Liu, X., Liu, D., Tan, C. and Feng, W. (2023) Gut microbiome-based machine learning for diagnostic prediction of liver fibrosis and cirrhosis: A systematic review and meta-analysis. BMC Medical Informatics and Decision Making, 23(1): 294. doi: 10.1186/s12911-023-02402-1.

Naik, S. N., Forlano, R., Manousou, P., Goldin, R. and Angelini, E. D. (2023) Fibrosis severity scoring on Sirius red histology with multiple-instance deep learning. Biological Imaging, 3. doi: 10.1017/s2633903x23000144.

Olianas, A. et al. (2023) Top-down proteomics detection of potential salivary biomarkers for autoimmune liver diseases classification. International Journal of Molecular Sciences, 24(2). doi: 10.3390/ijms24020959.

Pasta, A. et al. (2024) Hepatocellular carcinoma in patients with autoimmune hepatitis. Journal of Hepatology, 81(3): e131–e132. doi: 10.1016/j.jhep.2024.03.025.

Rokkas, T. et al. (2021) Comparative effectiveness of multiple different first-line treatment regimens for Helicobacter pylori infection: A network meta-analysis. Gastroenterology, 161(2): 495–507.e4. doi: 10.1053/j.gastro.2021.04.012.

Schwabl, P. et al. (2017) The FXR agonist PX20606 ameliorates portal hypertension by targeting vascular remodelling and sinusoidal dysfunction. Journal of Hepatology, 66(4): 724–733. doi: 10.1016/j.jhep.2016.12.005.

Terziroli Beretta-Piccoli, B., Mieli-Vergani, G. and Vergani, D. (2022) HLA, gut microbiome and hepatic autoimmunity. Frontiers in Immunology, 13. doi: 10.3389/fimmu.2022.980768.

Umapathy, L., Fu, Z., Philip, R., Martin, D., Altbach, M. and Bilgin, A. (2022) Learning to segment with limited

annotations: Self-supervised pretraining with regression and contrastive loss in MRI. Joint Annual Meeting ISMRM–ESMRMB, pp. 2–5. doi: 10.58530/2022/3914.

Wang, L. et al. (2024) IAIH-PG consensus for histological criteria of AIH: Multicentre validation with focus on chronic liver diseases in China. Liver International, 44(9): 2282–2292. doi: 10.1111/liv.15971.

Wei, Q. et al. (2025) An exploratory machine learning model for predicting advanced liver fibrosis in autoimmune hepatitis patients: A preliminary study. Annals of Hepatology, 30(1): 101754. doi: 10.1016/j.aohep.2024.101754.

Wu, M., He, X., Li, F., Zhu, J., Wang, S. and Burstein, P. D. (2023) Weakly supervised volumetric prostate registration for MRI–TRUS image driven by signed distance map. Computers in Biology and Medicine, 163: 107150. doi: 10.1016/j.compbiomed.2023.107150.