Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Phenotypic Heterogeneity in Ischemic Heart Disease Uncovered by Unsupervised Clustering of Routine Laboratory Parameters

Anam Ijaz

anamijaz2021@gmail.com

University of the Punjab

Sara Aslam

Superior University

Shabana NA

University of the Punjab

Research Article

Keywords: Ischemic heart disease, cardiovascular disease, unsupervised machine learning, patient phenotyping, hematological parameters, explainable AI

Posted Date: July 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10411130/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Ischemic heart disease (IHD) is the leading cause of global mortality. Whether routine, low-cost laboratory parameters can reveal clinically meaningful heterogeneity within an already-diagnosed IHD population remains unexplored.

Methods

In this cross-sectional study, 306 confirmed IHD patients from hospitals in Lahore, Pakistan, underwent measurement of routine hematological, hepatic, renal, and lipid parameters. An unsupervised machine learning framework was used to identify patient phenotypes directly from laboratory profiles. Data were split into training (70%) and test (30%) subsets; five clustering methods; k-means, Gaussian mixture models, fuzzy c-means, hierarchical clustering, and DBSCAN, were compared on UMAP-reduced features, and the best-generalizing configuration was selected. Random Forest, Decision Tree, and fuzzy rule-based classifiers were trained to identify discriminating biomarkers and translate clusters into clinician-readable rules.

Results

K-means (k = 3) gave the most stable partition (silhouette = 0.361 training, 0.370 test), yielding three phenotypes: a small (12.7%) atherogenic-hepatic-inflammatory cluster (elevated VLDL, triglycerides, transaminases, neutrophils); a larger (26.5%) pure dyslipidemia cluster; and a majority (60.8%) lower-risk cluster with the lowest atherogenic lipid burden. VLDL, total bilirubin, triglycerides, AST, ALT, and HDL were the principal discriminating features (Random Forest test accuracy 78.3%, AUC 0.926). Decision Tree and fuzzy classifiers translated these into interpretable rules (e.g., "IF VLDL is High AND AST is Low"), revealing an accuracy–interpretability trade-off.

Conclusions

Even without imaging, angiography, or omics data, routine laboratory parameters reveal reproducible phenotypic heterogeneity in IHD, centered on a lipid-metabolic and hepatic-inflammatory axis. This interpretable, low-cost framework may support scalable risk stratification in resource-limited settings, pending external and prognostic validation.

Introduction

Ischemic heart disease (IHD) remains the leading cause of mortality worldwide and continues to pose a major public health challenge. Cardiovascular diseases account for approximately 32% of all global

deaths, resulting in an estimated 19.8 million deaths annually, with the majority occurring in low- and middle-income countries[1]. Despite substantial advances in diagnostic techniques and treatment strategies, the burden of IHD remains disproportionately high, particularly in resource-limited healthcare settings where access to sophisticated diagnostic tools such as coronary angiography, cardiac imaging, and high-sensitivity biomarkers is often restricted. Consequently, early disease detection and effective risk stratification remain critical yet underexplored aspects of cardiovascular care.

The conventional diagnosis of IHD relies primarily on clinical symptoms, electrocardiographic findings, imaging modalities, and established biomarkers such as cardiac troponins and lipid profiles. However, these approaches are largely reactive and often detect the disease after significant pathophysiological changes have already occurred. Emerging evidence suggests that routinely measured laboratory parameters including hematological indices, renal function markers, liver enzymes, and electrolyte profiles, may reflect important biological processes such as systemic inflammation, endothelial dysfunction, oxidative stress, and coagulation abnormalities[2, 3]. These mechanisms play a central role in the initiation and progression of atherosclerosis and myocardial ischemia. However, their potential to reveal clinically meaningful subgroups, or phenotypes, among patients already diagnosed with IHD, rather than simply distinguishing IHD from non-IHD remains largely unexplored.

Recent advances in machine learning (ML) have transformed predictive medicine by enabling the analysis of large, complex, and multidimensional clinical datasets. Unlike traditional statistical approaches, ML algorithms can identify non-linear relationships and intricate interactions among variables, thereby improving prediction accuracy and clinical decision-making[4–6]. Existing studies have demonstrated that ML-based models often outperform conventional cardiovascular risk assessment tools by achieving superior accuracy, sensitivity, and specificity across diverse patient populations. Beyond outcome prediction, unsupervised ML techniques such as clustering have proven valuable for uncovering hidden patient subgroups, or phenotypes, within clinically heterogeneous diseases—an approach increasingly used to characterize disease heterogeneity in conditions such as sepsis, metabolic syndrome, and heart failure, but rarely applied to IHD using routine laboratory parameters[7].

Despite these advances, a significant translational gap remains in the clinical application of ML for IHD. Most existing predictive and stratification models depend on specialized data sources, including imaging studies, electrocardiographic signals, and omics-based biomarkers, which may not be readily available in routine healthcare practice. Moreover, most studies have been conducted in high-resource settings, limiting their applicability to low- and middle-income regions. There is particularly limited research focusing on routine laboratory investigations such as complete blood count (CBC), liver function tests (LFTs), renal profiles, and metabolic panels, despite their low cost, widespread availability, and ability to reflect overall physiological status[2].

In addition, current literature highlights several challenges associated with ML-based cardiovascular prediction, including model interpretability, data heterogeneity, overfitting, and inadequate external

validation. Addressing these limitations is essential for ensuring reliable clinical implementation and successful integration into real-world decision-support systems.

Therefore, the present study aims to characterize phenotypic heterogeneity among confirmed ischemic heart disease (IHD) patients using routine, low-cost hematological and biochemical parameters. Because no non-IHD comparison group was available, an unsupervised machine learning framework was adopted to identify clinically distinct patient subgroups directly from laboratory profiles, rather than to discriminate IHD from disease-free individuals. Specifically, the objectives are to: (i) identify subgroups (phenotypes) of IHD patients with distinct laboratory profiles using cluster analysis validated on a held-out test set, (ii) determine which routine biomarkers most strongly distinguish these phenotypes using feature importance ranking, and (iii) translate these findings into interpretable, clinician-readable rules using decision tree and fuzzy rule-based explainable artificial intelligence techniques[8]. By utilizing readily available clinical laboratory data, this study seeks to develop a scalable, cost-effective, and clinically interpretable framework for risk stratification within an already-diagnosed IHD population, applicable across diverse, resource-limited healthcare environments.

Ultimately, this research contributes to the growing field of precision medicine by demonstrating that routinely collected laboratory data can reveal clinically meaningful heterogeneity even within a single diagnostic category. The scope of this study is deliberately focused: it is a single-cohort, hypothesis-generating analysis restricted to patients with confirmed IHD, using only routinely available hematological and biochemical parameters, without imaging, angiographic, or omics-based data, and without a non-IHD comparison group or external validation cohort. Rather than proposing a diagnostic or prognostic tool, this work establishes proof-of-concept that unsupervised, explainable machine learning can stratify an already-diagnosed IHD population into biologically distinct, laboratory-defined phenotypes using inexpensive and universally accessible data. Identifying such phenotypes may support more individualized monitoring and management strategies for patients already living with ischemic heart disease, particularly in resource-limited settings where advanced diagnostic and prognostic tools remain inaccessible; confirming their clinical utility, however, will require prospective validation against hard outcomes in independent cohorts.

Methods

Study Design and Ethical Approval

This hospital-based cross-sectional study was conducted from February 2024 to July 2024. Data were collected at a single time point; because the data were cross-sectional, causality and temporal relationships cannot be inferred. Approval was granted by the Institute of Microbiology and Molecular Genetics Research Ethics and Biosafety Committee (Approval No. D/645/MMG). This committee serves as the institutional ethics review board for human subjects' research and oversees biosafety protocols. The study was performed in accordance with the Declaration of Helsinki.

Patient recruitment and data collection

A minimum sample size of approximately 217 participants was calculated using a single-proportion formula with finite population correction, assuming a 17% IHD prevalence, 95% confidence level, and 5% margin of error (Supplementary Information S1). Recruitment within the study period ultimately yielded confirmed IHD cases without a matched non-IHD comparison group; the present analysis is therefore restricted to characterizing heterogeneity within this case cohort. The study comprised 306 confirmed IHD cases recruited from various hospitals in Lahore, Pakistan. Diagnosis of IHD was confirmed by an on-duty cardiologist based on positive cardiac echocardiography, electrocardiogram (ECG) findings, clinical assessment, and serum cardiac biomarkers, including Troponin T/I. Only recently diagnosed patients who had not yet commenced lipid-lowering or antihypertensive medications were enrolled, to avoid the confounding effects of pharmacological treatment on lipid profiles. No non-IHD comparison (control) group was recruited; the analysis therefore characterizes phenotypic heterogeneity within the IHD case cohort itself, using an unsupervised analytical framework described in the Statistical Analysis section below.

Participants with pre-existing conditions, including liver disease, kidney disease, or malignancy, were excluded from the cohort. IHD patients with comorbid diabetes, hypertension, smoking history, and family history were included in the study, as these represent clinically relevant risk factors. Diabetes was defined as a fasting blood glucose (FBG) of ≥ 126 mg/dL or a 2-hour postprandial blood glucose of ≥ 200 mg/dL, in accordance with WHO diagnostic criteria. Hypertension was classified using the criteria defined in the Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure (JNC 7)[9]. Blood pressure was recorded as the means of two consecutive measurements taken after the participant had been seated and rested for five minutes. Anthropometric measurements, including height and weight, were recorded in the clinic by trained staff. Body mass index (BMI) was calculated as weight in kilograms divided by the square of height in meters (kg/m^2).

Demographic and clinical data were recorded through structured face-to-face interviews and review of participants' medical records. A standardized questionnaire was administered to collect information on age, gender, BMI, smoking status, and history of diabetes and hypertension. Hematological parameter reports including CBC, LFT, RFT and Lipid profile values were extracted from laboratory biochemical reports generated at the respective analysis centers.

Determination of hematological parameters

Hematological parameters were measured in standard units.

Complete Blood Count (CBC) test analysis:

Complete blood count parameters were measured using an automated hematology analyzer, with the following laboratory-defined normal reference ranges: white blood cell count (WBC, 4–11 ×10⁹/L), red blood cell count (RBC, 4.5–5.5 ×10¹²/L), hemoglobin (HGB, 13–17 g/dL), hematocrit (HCT, 40–50%), mean corpuscular volume (MCV, 76–96 fL), mean corpuscular hemoglobin (MCH, 27–32 pg), mean corpuscular hemoglobin concentration (MCHC, 31.5–34.5 g/dL), platelet count (PLT, 150–450 ×10⁹/L), neutrophils (%Neut, 40–80%), lymphocytes (%LYMP, 20–40%), and monocytes (%MONO, 2–10%). For certain parameters, patients' normal ranges are different according to their age group and gender.

Liver function tests (LFTs) analysis:

Liver function tests included alanine aminotransferase (ALT, U/L), aspartate aminotransferase (AST, U/L), total bilirubin (mg/dL), and alkaline phosphatase (U/L), measured using an automated chemistry analyzer, with the following laboratory-defined normal reference ranges: ALT, 10–41 U/L; AST, 10–50 U/L; total bilirubin, up to 1.0 mg/dL; and alkaline phosphatase, 40–129 U/L.

For all liver function assays, different groups were made based on their normal value ranges

Renal function tests (RFTs) analysis:

Renal function tests included serum urea (mg/dL), creatinine (mg/dL), sodium (mmol/L), potassium (mmol/L), and chloride (mmol/L), measured using an automated chemistry analyzer, with the following laboratory-defined normal reference ranges: urea, 15–45 mg/dL; creatinine, 0.7–1.4 mg/dL; sodium, 137–145 mmol/L; potassium, 3.3–5.1 mmol/L; and chloride, 95–108 mmol/L. For all renal function assays, patients were additionally grouped on the basis of established normal reference ranges.

Blood Sample Collection

All participants were instructed to fast for at least 10–12 hours overnight prior to blood collection. Venous blood samples were collected in sterile vacutainer tubes under aseptic conditions and allowed to clot at room temperature for 10–20 minutes. Serum was separated by centrifugation at 3000 rpm for 5 min, and transferred to labelled Eppendorf tubes, and stored at − 20°C until biochemical analysis.

Prior to biochemical analysis, all samples were prescreened for human immunodeficiency virus (HIV), hepatitis B virus (HBV), hepatitis C virus (HCV), and syphilis, and reactive samples were excluded.

Statistical analysis

Machine Learning Models

All machine learning (ML) analyses were implemented in Python (version 3.10.7) using Pandas and NumPy for data handling; scikit-learn for preprocessing, clustering, dimensionality-reduction

comparison, Random Forest[10], and Decision Tree models[11]; UMAP-learn for nonlinear dimensionality reduction prior to clustering; scikit-fuzzy and ex-fuzzy for fuzzy c-means clustering and fuzzy rule-based classification[12]; SciPy for statistical testing; and Matplotlib, Seaborn, and Plotly for visualization. No deep learning models were used in the final analysis. The complete analytical workflow is illustrated in Supplementary Figure S1.

Dataset Partitioning and Preprocessing

Prior to analysis, variables with more than 5% missing values across the cohort were excluded from quantitative analysis—leaving 30 variables (26 continuous laboratory parameters and four binary clinical variables: gender, diabetes, hypertension, and family history) for downstream analysis.

The cohort was then split into training (70%, n = 214) and test (30%, n = 92) subsets using a random, non-stratified split (random_state = 42). This split was performed before any imputation, outlier handling, or scaling, and all subsequent preprocessing parameters were fit exclusively on the training subset and applied to the test subset, to prevent information from the held-out test set from influencing how the training data were cleaned or scaled.

Missing values in continuous variables were imputed using a skewness-aware strategy: for each variable, skewness was computed from the training subset only; if |skewness| > 1, missing values were imputed with the training-set median, otherwise with the training-set mean. The same training-derived imputation values were then applied to the test subset. Extreme outlier values were winsorised at five times the interquartile range (Q1 − 5×IQR, Q3 + 5×IQR), with bounds likewise computed on the training subset only and applied to both subsets; this conservative threshold was intended to correct likely data-entry errors while preserving genuine biological extremes. All continuous variables were subsequently scaled to a 0–1 range using min-max normalization, with the scaler fitted on the training subset and applied to both subsets.

Gender, diabetes status, hypertension status, and family history are inherently binary (two-level) variables and were represented as single binary (0/1) indicator columns (smoking status was already excluded at the missingness-filtering step described above); one-hot encoding with reference-level dropping applied to a two-level variable is mathematically equivalent to direct binary encoding and did not expand the feature space beyond the original four indicator columns.

Missing values in these four indicator columns (2.9–5.5% each) were imputed using mode substitution — the most frequently occurring category, computed on the training subset only and applied to both the training and test subsets, consistent with the leakage-safe approach used for the continuous variables. These binary indicators were used only in the descriptive and inferential analyses (Mann-Whitney U group comparisons, chi-square tests of categorical association, and PCA visualization); they were not included as input features in the unsupervised clustering, Random Forest, Decision Tree, or fuzzy rule-based models, which were trained exclusively on the 26 continuous laboratory parameters. Keeping

categorical risk-factor flags separate from the continuous biomarker feature space avoided mixing variable types within the distance-based clustering step.

Exploratory and Inferential Statistical Methods

Normality of continuous variables was assessed using the Shapiro–Wilk test (samples capped at 200 observations per variable for computational stability); only one of 26 continuous variables (Age) was normally distributed ($p > 0.05$), so non-parametric methods were used throughout. Pairwise associations among continuous variables were assessed using Spearman's rank correlation. Differences in key biomarkers across binary clinical groups (gender, diabetes, hypertension, family history) were assessed using the Mann–Whitney U test. Associations among categorical variables were assessed using the chi-square test of independence, with Cramér's V as a measure of effect size. Principal component analysis (PCA) was used descriptively to characterize overall variance structure across the full cohort. Outlying observations were flagged using both a univariate z-score threshold ($|z|>3$) and a multivariate Isolation Forest (contamination = 5%). A p-value < 0.05 was considered statistically significant throughout.

Unsupervised Clustering and Cross-Validation

Because no non-IHD comparison group was available, patient phenotypes were identified using unsupervised clustering rather than supervised classification. The 26 normalized continuous variables were first reduced to five dimensions using Uniform Manifold Approximation and Projection (UMAP; fitted on the training subset, with the same mapping applied to the test subset) to reduce noise prior to clustering. Five clustering paradigms were compared on the training subset: k-means and Gaussian mixture models ($k = 3$ and $k = 4$), fuzzy c-means[13] ($k = 3$ and $k = 4$), Ward-linkage hierarchical clustering[14] ($k = 3$ and $k = 4$), and density-based spatial clustering[15] (DBSCAN, with eps and min_samples tuned to yield 3–4 clusters with < 30% of points classified as noise). The configuration with the highest training-set silhouette score was provisionally selected; each test patient was then assigned to the nearest training-set cluster centroid (Euclidean distance in UMAP space), and a test-set silhouette score was computed to assess whether the cluster structure generalized to unseen patients. The clustering solution reported is the configuration with the best combination of training- and test-set silhouette scores, i.e. the smallest train–test gap alongside a high absolute score.

Feature Importance and Interpretable Rule Extraction

To identify which laboratory parameters most strongly distinguished the resulting clusters, a Random Forest classifier (500 trees, minimum 5 samples per leaf, balanced class weights) was trained on the training subset to predict cluster membership from the original, non-UMAP-reduced variables, with the minimal feature subset selected via recursive feature elimination with 5-fold stratified cross-validation (RFECV). A shallow Decision Tree (maximum depth 4, minimum 10 samples per leaf) was additionally trained on the same features to produce fully transparent, clinician-readable splitting rules. Finally, an evolutionary fuzzy rule-based classifier (ex-fuzzy package) was trained on the top seven Random-Forest-ranked features, using three linguistic terms per variable (Low/Medium/High) and a maximum of 20 IF–

THEN rules, to generate interpretable fuzzy decision rules describing each phenotype. All three interpretable models were trained on the training subset only and evaluated on the held-out test subset.

Machine learning analysis and model performance parameters

Clustering quality was quantified using the silhouette coefficient, computed separately on the training and test subsets, with the train–test difference (gap) used to assess generalizability (gap < 0.05 classified as excellent, < 0.15 as acceptable, ≥ 0.15 as indicative of overfitting). Classification performance of the Random Forest and Decision Tree models in predicting cluster membership was quantified using accuracy, precision, recall, and F1-score (per cluster and as macro- and weighted averages) on both the training and held-out test subsets, with confusion matrices used to visualize misclassification patterns. Fuzzy rule quality was quantified using rule-level accuracy (ACC, the proportion of covered patients correctly classified by that rule) and degree of support (DS, the proportion of patients to which a given rule applied), in addition to overall train and test classification accuracy of the complete fuzzy rule set.

Results

Cohort Characteristics and Subgroup Biomarker Analysis

The cohort comprised 306 secondary-care patients (60.1% male), with high prevalences of hypertension (65.0%) and diabetes (56.5%), and positive family history in 56.5% (Fig. 1A). Age distributions were unimodal, centered in the sixth decade for both sexes (Fig. 1B). The secondary-care composition limits generalizability to unselected community populations. Chi-square analysis identified significant pairwise associations between diabetes–hypertension ($\chi^2 = 9.89$, $p = 0.002$, Cramér's V = 0.18), gender–hypertension ($\chi^2 = 6.19$, $p = 0.013$, Cramér's V = 0.14), and diabetes–family history ($\chi^2 = 4.09$, $p = 0.043$, Cramér's V = 0.12) (Supplementary Table S1); however, all effect sizes were small (Cramér's V < 0.20), and three pairs were non-significant, confirming that binary clinical labels share insufficient variance to serve as proxies for one another and motivating the multivariate unsupervised approach adopted here.

Sex stratification revealed significantly higher hemoglobin (HGB) in males ($p < 0.001$, Fig. 1C), consistent with androgen-mediated erythropoiesis; this sex differential in hematological indices represents a potential confound in unsupervised analyses if HGB and co-varying erythrocyte parameters are not considered in the context of sex composition. Among diabetic patients, HDL cholesterol was significantly lower ($p = 0.011$) and WBC count modestly elevated ($p = 0.047$) compared with non-diabetic counterparts (Fig. 1D), consistent with insulin-resistance-associated dyslipidemias and systemic low-grade inflammation respectively. No biomarker achieved significance between hypertensive and normotensive groups across the full panel (all $p > 0.05$, Fig. 1E), a null result most plausibly attributable to pharmacological attenuation — antihypertensive and statin use being prevalent yet unrecorded in this dataset — rather than true biological equivalence. Collectively, subgroup analysis confirms that

established diagnostic labels capture only a fraction of continuous biomarker variation, justifying a data-driven stratification framework.

Correlation Structure and Dimensionality Reduction

Spearman correlation analysis across 20 standardized biomarkers (Fig. 2; Supplementary Table S2) revealed strong intra-domain coupling: HCT–HGB (ρ = 0.86), MCH–MCV (ρ = 0.81), and %LYMP–%Neut (ρ = −0.81, a mathematical artefact of the differential sum constraint rather than an independent biological signal). Renal (creatinine–urea, ρ = 0.62) and hepatic (ALT–AST, ρ = 0.57) pairs showed moderate concordance. Cross-domain correlations were universally weak (|ρ| < 0.30), confirming relative orthogonality across hematological, lipid, and renal compartments and validating the premise that multivariate clustering would capture complementary rather than redundant biological axes.

PCA explained 13.9% and 10.0% variance on PC1 and PC2 respectively, requiring 12 components to reach 80% cumulative variance (Fig. 3A), indicating moderate intrinsic dimensionality with no single dominant axis. Two-dimensional projections colored by all four clinical labels (diabetes, hypertension, sex, family history) showed substantial inter-group overlap (Fig. 3B), confirming that pre-existing diagnoses do not align with the principal axes of continuous biomarker variation. Isolation Forest anomaly detection identified approximately 15–20 peripheral patients in PCA space (Fig. 3C); these were retained in the analysis as potential high-risk phenotypic extremes. Univariate z-score screening (|z| > 3) identified the highest outlier burden in electrolytes (chloride n ≈ 23, sodium n ≈ 19) and hepatic/renal markers (Fig. 3D), warranting cautious interpretation of cluster boundaries involving these variables.

Clustering Method Comparison and Selection

Nine configurations spanning K-Means, DBSCAN, GMM, FCM, and hierarchical clustering at k = 3 and k = 4 were evaluated by Silhouette coefficient on stratified 70/30 train–test splits (Table 1, Fig. 4A). DBSCAN achieved the highest training Silhouette (0.393) but offers limited clinical interpretability and hyperparameters without biological analogues. K-Means_k3 produced train and test Silhouette scores of 0.361 and 0.370 respectively — with the test score marginally exceeding training, a pattern indicating robust, generalizable cluster geometry. All k = 4 configurations underperformed their k = 3 counterparts across all algorithm families and fell at or below the 0.30 acceptability threshold, constituting convergent evidence that a three-cluster partition is the more parsimonious solution for this dataset. No method exceeded the 0.15 generalization gap overfit threshold (Fig. 4B); K-Means_k3 and GMM_k3 produced negative gaps (− 0.009 and − 0.007 respectively), further confirming structural stability. Based on test Silhouette performance, minimal generalization gap, and interpretability, K-Means_k3 was selected as the reference partition.

Table 1
All clustering methods ranked by train silhouette in 70:30 datasets

Method	k	Train Silhouette	Test Silhouette
KMeans_k3	3	0.073	0.360
Hierarchical_k3	3	0.346	0.360
DBSCAN	3	0.373	0.356
KMeans_k4	4	0.074	0.321
Hierarchical_k4	4	0.316	0.313
GMM_k3	3	0.252	0.297
GMM_k4	4	0.324	0.295
FCM_k3	3	0.256	0.289
FCM_k4	4	0.270	0.273

Cluster Phenotype Characterization

K-Means_k3 partitioned the cohort into Cluster 1 (n = 186, 60.8%), Cluster 2 (n = 81, 26.5%), and Cluster 0 (n = 39, 12.7%), with consistent centroid positions across training and test PCA projections confirming spatial stability (Fig. 5A). Z-score biomarker profiles and radar charts characterized each phenotype relative to the cohort mean (Fig. 5B–C).

Cluster 0 (12.7%) was defined by co-elevation of VLDL (z = + 1.12), TG (z = + 1.14 mg/dL), AST (z = + 1.15), ALT (z = + 1.15), WBC (z = + 0.80 ×10³), and %Neut (z = + 1.15), with low % lymphocytes — a combined atherogenic, hepatic, and pro-inflammatory signature consistent with metabolic-associated steatotic liver disease and systemic inflammation. Cluster 2 (26.5%) exhibited a pure dyslipidemia profile: elevated VLDL (z = + 1.12), TG (z = + 1.14), and LDL (z = + 1.10) without transaminase elevation, representing a classical mixed hyperlipidemia phenotype. Cluster 1 (60.8%) showed the lowest atherogenic lipid burden (TG z = − 0.71, VLDL z = − 0.86 mg/dL) and the highest %lymphocytes, consistent with a comparatively lower inflammatory and dyslipidemia risk profile. These phenotypic labels are hypothesis-generating interpretations grounded in pathophysiology; without external validation against clinical outcomes, they should not be construed as confirmed clinical entities.

Cluster Classification and Interpretable Rule Extraction

A Random Forest (RF) classifier with Recursive Feature Elimination identified VLDL (importance ≈ 0.175), total bilirubin (≈ 0.145 mg/dL), TG (≈ 0.130 mg/dL), AST, ALT, and HDL as the principal discriminative features (Fig. 5D), with hematological and renal markers contributing negligibly — corroborating the cluster profiling finding that lipid metabolism and hepatic function are the dominant inter-cluster axes.

The RF achieved training accuracy of 93.9% (weighted F1 = 0.94, Table 2) and test accuracy of 78.3% (weighted F1 = 0.80, Table 3), with test-set micro-average AUC of 0.926 (Fig. 5E-F). Cluster 0's test precision collapsed to 0.43 despite perfect recall (1.00), a consequence of class imbalance (n = 10 on test) and biomarker boundary overlap with Cluster 2; the absence of oversampling or cost-sensitive correction is a methodological limitation that future work should address given the clinical importance of correctly identifying this high-risk phenotype.

Table 2
Random forest train Set classification report

Cluster	Precision	Recall	F1-Score	Support
Cluster 0	0.96	0.97	0.96	68
Cluster 1	0.99	0.98	0.98	123
Cluster 2	0.96	1.00	0.98	23
Macro avg	0.97	0.98	0.98	214
Weighted avg	0.98	0.98	0.98	214

Table 3
Random forest test Set classification report.

Cluster	Precision	Recall	F1-Score	Support
Cluster 0	0.82	0.87	0.84	31
Cluster 1	0.94	0.89	0.91	53
Cluster 2	0.70	0.78	0.74	9
Macro avg	0.82	0.85	0.83	93
Weighted avg	0.88	0.87	0.87	93

A Decision Tree (DT) classifier provided interpretable assignment rules, partitioning patients first by AST, then VLDL and ALT, with deeper nodes involving total bilirubin, TG, HGB, and urea (Fig. 6A) a hierarchy concordant with RF feature importance and independently reinforcing hepatic and lipid biomarkers as primary discriminators. The DT achieved training accuracy of 83.6% but test accuracy of only 59.8% (weighted F1 = 0.61, Tables 4–5), a 23.8-percentage-point gap attributable to greedy partitioning overfitting, unstable small-class leaf statistics (n = 10 for Cluster 0 on test), and biologically marginal cut-points in secondary splits. This degradation is a classifier property, not evidence against cluster validity, as the RF achieved 78.3% accuracy on identical data. Test ROC AUCs for the DT ranged from 0.875 (Cluster 0) to 0.642 (Cluster 2) (Fig. 6B-C), confirming the greatest generalization failure for the intermediate dyslipidemia phenotype whose biomarker boundaries with Cluster 1 are inherently less crisp.

Table 4
Decision Tree train Set classification report.

Cluster	Precision	Recall	F1-Score	Support
Cluster 0	0.63	0.90	0.74	68
Cluster 1	0.96	0.67	0.79	123
Cluster 2	0.72	1.00	0.84	23
Macro avg	0.77	0.85	0.79	214
Weighted avg	0.83	0.78	0.78	214

Table 5
Decision Tree test Set classification report.

Cluster	Precision	Recall	F1-Score	Support
Cluster 0	0.56	0.90	0.69	31
Cluster 1	0.88	0.53	0.66	53
Cluster 2	0.73	0.89	0.80	9
Macro avg	0.72	0.77	0.72	93
Weighted avg	0.76	0.69	0.68	93

The ExFuzzy classifier extracted eight linguistic rules, achieving consistent training and test accuracy (≈ 44% in both partitions, Fig. 7A), confirming that its performance ceiling is set by linguistic variable discretization rather than overfitting. High-accuracy Cluster 2 rules, 'IF VLDL IS High AND AST IS Low' (≈ 100%) and 'IF VLDL IS High AND LDL IS Medium' (≈ 95%), demonstrate that the atherogenic phenotype is parsimoniously described by two routine laboratory parameters. Cluster 1's dominant rule ('IF VLDL IS Low AND HDL IS Low') achieved the highest patient coverage across all rules (Fig. 7B-C), efficiently routing the majority cluster. Cluster 0 rules achieved accuracy below 50% and minimal coverage, reflecting the phenotypic complexity and small size of this group. Across all three classifiers, an accuracy–interpretability trade-off is evident: RF maximizes predictive performance; the DT provides transparent clinical rules at the cost of generalization; and ExFuzzy offers the most linguistically accessible and overfit-resistant framework at lower absolute accuracy. These approaches are complementary rather than competing, suggesting a staged implementation where RF performs population-level screening, the DT enables rapid triage, and ExFuzzy supports patient-level explainable decision support.

Discussion

This study demonstrates that clinically meaningful phenotypic heterogeneity exists among confirmed ischemic heart disease (IHD) patients even when derived exclusively from routinely collected, low-cost

hematological and biochemical parameters, without recourse to coronary imaging, angiography, or omics-based platforms. The three-cluster K-Means solution, validated by a minimal train–test silhouette gap, resolved the cohort into a small, high-risk atherogenic-hepatic-inflammatory phenotype (Cluster 0), a larger pure dyslipidemia phenotype (Cluster 2), and a majority lower-risk phenotype (Cluster 1) with the lowest atherogenic lipid burden and the highest lymphocyte proportion. Three independent interpretable classifiers — Random Forest, Decision Tree, and an evolutionary fuzzy rule-based system — converged on the same underlying biological axis, namely VLDL, triglycerides, transaminases, and total bilirubin, as the principal discriminators of cluster membership, indicating that lipid metabolism and hepatic function, rather than classical hematological or renal indices, delineate phenotypic subgroups within an already-diagnosed IHD population.

Our findings extend, rather than replicate, prior unsupervised machine learning work in cardiovascular disease. Hu et al. applied K-means clustering to a large electronic-medical-record cohort of 155,894 patients to discriminate cardiovascular disease from non-cardiovascular-disease cases using 2-, 4-, and 8-cluster solutions combined with a Bayesian-theorem-based predictive framework, achieving classification accuracies of approximately 85–87%[16]. This task differs fundamentally from ours: whereas their clustering separated diseased from non-diseased individuals, the present study characterized heterogeneity within a single, already-diagnosed IHD population, since no non-IHD comparator group was available. Both studies nonetheless converge on the principle that unsupervised partitioning of routinely available laboratory or record-based data recovers clinically coherent structure without requiring specialized inputs, supporting the broader feasibility of laboratory-based phenotyping in settings where imaging or procedural data are unavailable.

A conceptually closer comparator is the hierarchical-clustering analysis of 3205 STEMI patients by Birdal et al., who partitioned patients into two phenotypic clusters using a combination of clinical, laboratory, and angiographic variables and demonstrated divergent short-term outcomes between groups[17]. Although our cohort was more granular, resolving into three clusters, and was restricted to laboratory variables alone, both studies support the premise that unsupervised clustering of readily available admission data recovers prognostically distinct IHD phenotypes. The absence of angiographic and outcome data in our dataset, however, precluded direct outcome validation of the clusters identified here, a limitation shared with many data-driven phenotyping studies performed on single, cross-sectional cohorts.

The identification of a small, biologically distinct high-risk subgroup — Cluster 0, comprising 12.7% of the cohort — parallels findings from a recent unsupervised clustering study of myocardial infarction complicated by out-of-hospital cardiac arrest, in which k-means clustering of 478 patients identified a minority cardiogenic-shock phenotype (21% of the cohort) characterized by high lactate, reduced ejection fraction, and markedly elevated 90-day mortality (77.2% versus 22.5% in the lowest-risk cluster) [18]. While the present study lacks outcome data to confirm a comparable mortality gradient, the recurrence of a small, severely deranged subgroup embedded within a larger, more homogeneous population is a consistent structural feature across cardiovascular phenotyping studies, consistent with

the hypothesis that Cluster 0 — defined by combined dyslipidemia, transaminitis, and neutrophilia — represents a metabolically and inflammatorily distinct high-risk phenotype warranting prioritized clinical attention and, ideally, prospective outcome validation.

Methodologically, our comparison of five unsupervised algorithms — k-means, Gaussian mixture models, fuzzy c-means, hierarchical clustering, and DBSCAN — and selection of the configuration with the best train–test silhouette generalization mirrors the rigorous multi-method validation strategy employed by Banerjee et al. in their analysis of over 320,000 heart failure patients across three electronic health record sources, which combined four clustering methods with external, prognostic, and genetic validation to identify five reproducible heart failure subtypes[19]. The scale and validation depth of that study — spanning independent cohorts and polygenic risk scores — substantially exceeds what is feasible in a single-center cohort of 306 patients, and represents an important direction for future work: external replication of the present three-cluster solution in an independent IHD cohort, together with prognostic follow-up, would be required before these phenotypes could be considered validated clinical entities rather than hypothesis-generating groupings.

The biological axis identified here — combined dyslipidemia and low-grade inflammation — also resonates with recent proteomic phenotyping work in coronary artery disease. Distinct immune-metabolic phenotypes have similarly been described in patients with poor coronary collateral circulation, where two reproducible subtypes — one driven by vascular remodeling and one by immuno-thrombotic myocardial dysfunction — were distinguished by inflammatory cell counts, eosinophil/basophil proportions, and metabolic markers, with the immuno-thrombotic phenotype carrying the highest incidence of major adverse cardiovascular events on external validation[20]. That distinct immune- and lipid-driven axes recur across two very different analytic platforms — circulating proteomics in that study and routine hematological/biochemical indices in ours — lends indirect support to the biological plausibility of the phenotypes identified here, even though our findings await outcome-based confirmation.

Consistent with the wider unsupervised-phenotyping literature in coronary artery disease, Williams et al. identified three clusters among 9221 patients with known coronary artery disease undergoing SPECT myocardial perfusion imaging and showed that cluster membership predicted all-cause mortality more accurately than ischemic burden alone [21]. The recurrence of a three-cluster structure across strikingly different data modalities — clinical, acquisition, and imaging parameters in that cohort versus routine laboratory parameters here — suggests that a three-phenotype partition may be a relatively stable feature of heterogeneous IHD/CAD populations, although this observation should be interpreted cautiously given differences in sample size, case-mix, and clustering methodology between studies.

A further point of comparison concerns the molecular pathways nominally implicated by omics-based studies. Multi-omics prediction of cardiovascular disease risk factors from metabolomic, epigenetic, and transcriptomic data has shown that multi-omics integration generally outperforms single-omics prediction, particularly for blood-pressure-related outcomes[22], while targeted metabolomic profiling

combined with machine learning has identified panels of circulating metabolites — including markers of lipid and amino-acid metabolism — capable of diagnosing acute myocardial ischemia with high accuracy[23]. Similarly, bioinformatic and machine-learning analyses of gene-expression data in cardiomyopathy have implicated suppressed immune-associated genes and hub genes such as HSPA8 in disease pathogenesis [24]. Although the present study did not employ omics platforms, the convergence of our Random-Forest-ranked features — VLDL, total bilirubin, triglycerides, and transaminases — on a combined lipid-metabolic and hepatic-inflammatory axis is broadly congruent with the metabolic and immune pathways nominated by these molecular studies, raising the possibility that routine, low-cost laboratory panels may serve as accessible clinical surrogates for at least part of the same underlying pathobiology that more expensive multi-omics platforms are designed to detect — an attractive proposition for resource-limited settings such as ours, where metabolomic or transcriptomic profiling is rarely feasible.

None of the comparator studies reviewed here report fully interpretable rule extraction from their cluster solutions; most rely on post hoc descriptive comparison of cluster characteristics[17, 18, 21] or supervised prediction models trained on omics features[22–24] without translating cluster assignment into transparent, clinician-readable decision rules. The staged interpretability framework used in the present study — Random Forest for population-level ranking, a shallow Decision Tree for rapid bedside triage, and an evolutionary fuzzy classifier for linguistically transparent, patient-level explanation — addresses a translational gap evident across this literature: unsupervised phenotyping, however clinically informative, rarely produces outputs that a clinician without machine-learning expertise can act upon directly. By contrast, rules such as ‘IF VLDL is High AND AST is Low’ offer an immediately interpretable heuristic for flagging the atherogenic-hepatic phenotypes using only two routine laboratory values, illustrating how explainable AI can translate cluster-level insight into bedside-usable decision support — a feature particularly relevant in resource-limited settings where black-box models are unlikely to be trusted or adopted.

Collectively, these comparisons situate the present findings within a rapidly growing but methodologically heterogeneous body of literature applying unsupervised machine learning to cardiovascular phenotyping — spanning electronic health records[16, 19], cardiac arrest and infarction registries[17, 18], proteomic and metabolomic platforms[22, 23], gene-expression datasets[24], and multimodal imaging[21]. What distinguishes the present study is not analytic novelty per se, but the demonstration that comparable phenotypic structure — a small high-risk inflammatory-metabolic cluster, an intermediate dyslipidemia cluster, and a larger lower-risk cluster — can be recovered from inexpensive, universally available laboratory panels in a modestly resourced, single-center South Asian cohort, without recourse to angiography, cardiac imaging, or the omics platforms used in most comparator studies. This has direct relevance for cardiovascular risk stratification in low- and middle-income countries, where the infrastructure required for imaging- or omics-based phenotyping is frequently unavailable, but complete blood counts, liver and renal function tests, and lipid panels are routinely collected as part of standard IHD work-up. Whether the phenotypes identified here translate

into differential clinical outcomes remains an open question, highlighting the need for future prospective studies to validate these findings.

Conclusion

This study shows that unsupervised clustering of routinely available hematological and biochemical parameters can resolve clinically coherent phenotypic heterogeneity among confirmed IHD patients, even without coronary imaging, angiography, or omics-based platforms. Three reproducible phenotypes emerged, distinguished chiefly by a combined lipid-metabolic and hepatic-inflammatory axis (VLDL, triglycerides, transaminases, and total bilirubin) and were successfully translated into transparent, clinician-readable decision rules via Random Forest, Decision Tree, and evolutionary fuzzy classifiers. These findings suggest that inexpensive, universally available laboratory panels can serve as an accessible alternative or complement to imaging- and omics-based phenotyping approaches, particularly in low- and middle-income healthcare settings where such infrastructure is scarce. Prospective studies incorporating clinical outcomes, external validation across independent cohorts, and comparison against non-IHD populations will be required to confirm the prognostic significance of these phenotypes and to establish their utility for individualized risk stratification and monitoring in ischemic heart disease.

Declarations

Ethics approval and consent to participate: The research was approved from the Research Ethics and Biosafety Committee of the University of the Punjab, Pakistan which follows the principles of Declaration of Helsinki. Written informed consent was obtained from all participants prior to enrollment.

Consent for publication: Not applicable.

Availability of data and materials: The dataset analyzed during this study is available as Supplementary Data S1. The complete analysis code is publicly accessible via the GitHub repository: https://github.com/anamijaz2021-qmul/Hematological-parameters-with-exfuzzy

Competing interests: The authors have no competing interests.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions: A.I.: Designed the study, data collection, data curation and analysis, wrote and proofread the manuscript. **S.A.:** Data analysis, prepared tables, figures, wrote and proofread the manuscript. **Sh.:** Proofread, reviewed, and approved the draft of manuscript.

Acknowledgements: not applicable

References

1. Organization WH. World Health Organization cardiovascular diseases (CVDs) fact sheet. World Health Organ. 2020;42(1):207–16.
2. Dhingra R, Vasan RS. Biomarkers in cardiovascular disease: Statistical assessment and section on key novel heart failure biomarkers. Trends Cardiovasc Med. 2017;27(2):123–33.
3. Popescu D, et al. Cardiovascular risk assessment in rheumatoid arthritis: accelerated atherosclerosis, new biomarkers, and the effects of biological therapy. Life. 2023;13(2):319.
4. Wu J. Artificial intelligence in cardiovascular medicine. Curr Cardiol Rev. 2021;17(3):157–67.
5. Singhal AA, et al. Clinical Applications of Artificial Intelligence in Cardiovascular Imaging: Where Do We Stand? Life. 2026;16(3):507.
6. Rajpurkar P, et al. AI in health and medicine. Nat Med. 2022;28(1):31–8.
7. Banerjee T, Paçal İ. A systematic review of machine learning in heart disease prediction. Turkish J Biology. 2025;49(5):600–34.
8. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst, 2017. 30.
9. Chobanian AV, et al. Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure. Hypertension. 2003;42(6):1206–52.
10. Parmar A, Katariya R, Patel V. *A review on random forest: An ensemble classifier*. in *International conference on intelligent data communication technologies and internet of things*. 2018. Springer.
11. Priyanka, Kumar D. Decision tree classifier: a detailed survey. Int J Inform Decis Sci. 2020;12(3):246–69.
12. Fumanal-Idocin J, Andreu-Perez J. Ex-Fuzzy: A library for symbolic explainable AI through fuzzy logic programming. Neurocomputing. 2024;599:128048.
13. Hashemi SE, Gholian-Jouybari F, Hajiaghaei-Keshteli M. A fuzzy C-means algorithm for optimizing data clustering. Expert Syst Appl. 2023;227:120377.
14. Yazdi F, Asadi S. Enhancing cardiovascular disease diagnosis: The power of optimized ensemble learning. IEEE Access. 2025;13:46747–62.
15. Campello RJ et al. *Density-based clustering.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2020. 10(2): p. e1343.
16. Hu Y, et al. Detecting cardiovascular diseases using unsupervised machine learning clustering based on electronic medical records. BMC Med Res Methodol. 2024;24(1):309.
17. Birdal O, et al. Cluster analysis of clinical, angiographic, and laboratory parameters in patients with ST-segment elevation myocardial infarction. Lipids Health Dis. 2024;23(1):166.
18. Singh M et al. *Phenotypic clustering of myocardial infarction complicated by out-of-hospital cardiac arrest using unsupervised machine learning.* Resuscitation, 2026: p. 111037.

19. Banerjee A, et al. Identifying subtypes of heart failure from three electronic health record sources with machine learning: an external, prognostic, and genetic validation study. Lancet Digit Health. 2023;5(6):e370–9.
20. Guo Z-T et al. Distinct immune-metabolic phenotypes underlie poor coronary collateral circulation. Cardiovasc Diabetol, 2025. 24(1).
21. Williams MC, et al. Unsupervised learning to characterize patients with known coronary artery disease undergoing myocardial perfusion imaging. Eur J Nucl Med Mol Imaging. 2023;50(9):2656–68.
22. Drouard G, et al. Exploring machine learning strategies for predicting cardiovascular disease risk factors from multi-omic data. BMC Med Inf Decis Mak. 2024;24(1):116.
23. Cao J, et al. Combined metabolomics and machine learning algorithms to explore metabolic biomarkers for diagnosis of acute myocardial ischemia. Int J Legal Med. 2023;137(1):169–80.
24. Wang W, et al. Potential diagnostic biomarkers in heart failure: Suppressed immune-associated genes identified by bioinformatic analysis and machine learning. Eur J Pharmacol. 2025;986:177153.

Figures

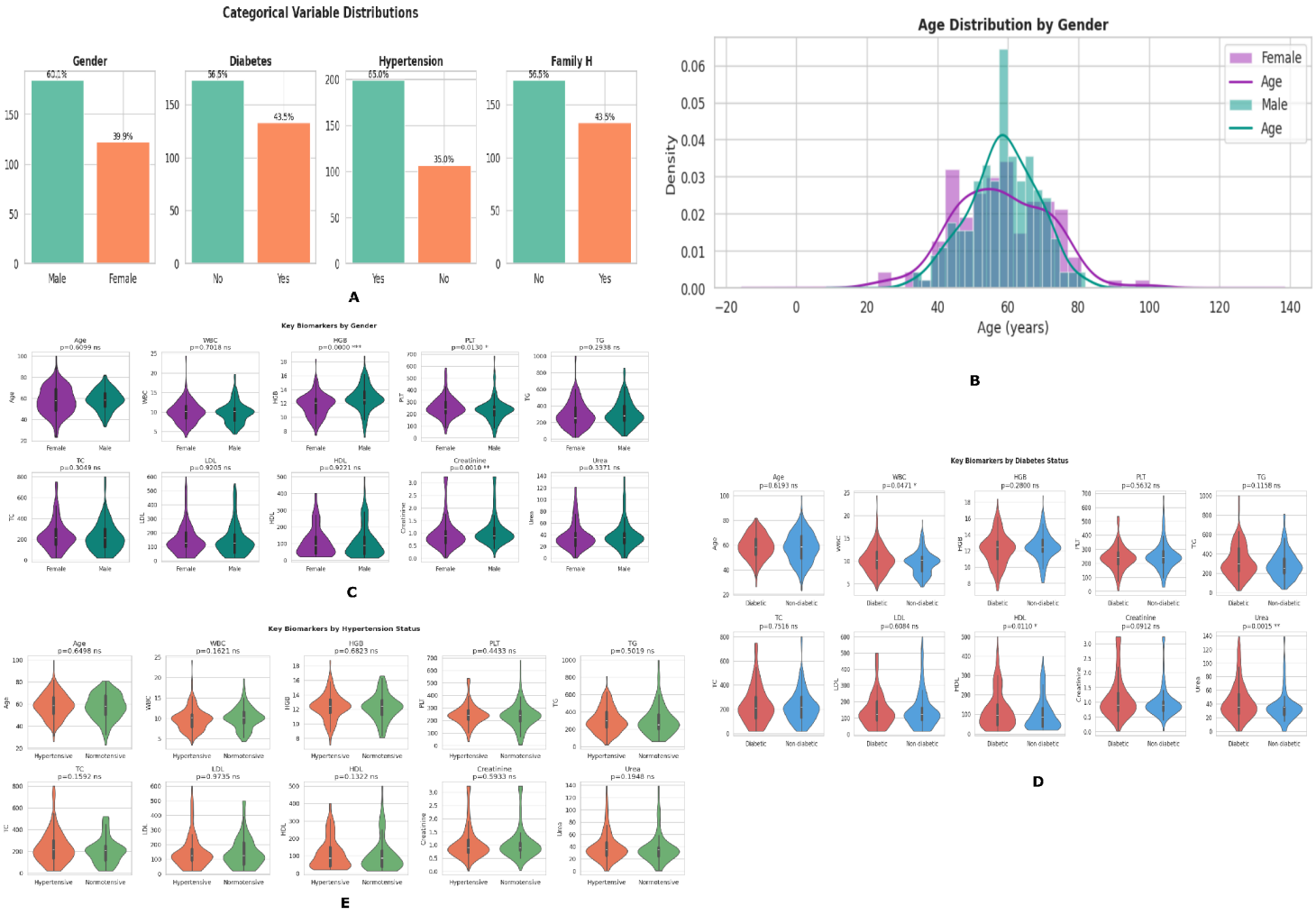


Figure 1

Descriptive exploratory data analysis of the patient cohort

(A) Distribution of categorical variables: sex, diabetes status, hypertension status, and family history, with percentages of each category. (B) Age distribution of the cohort stratified by gender, shown as overlaid histograms with fitted kernel density curves. (C) Violin plots comparing key biomarkers (age, WBC, HGB, PLT, TG, TC, LDL, HDL, creatinine, urea) between female and male patients. (D) Violin plots comparing the same biomarkers between diabetic and non-diabetic patients. (E) Violin plots comparing the same biomarkers between hypertensive and normotensive patients. Asterisks denote significance level (*p < 0.05, **p < 0.01, ***p < 0.001; ns = not significant).

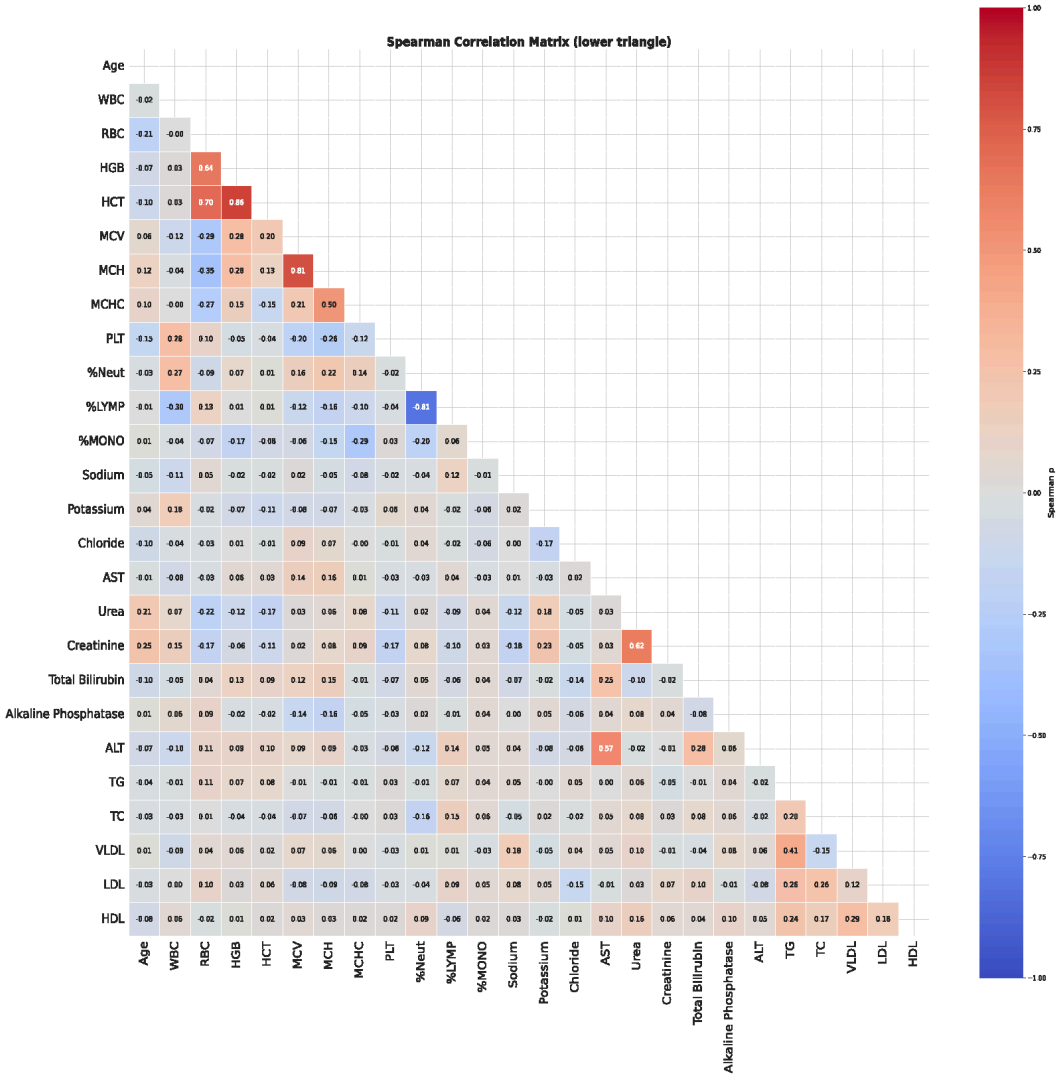


Figure 2

Spearman correlation matrix of clinical and biochemical variables

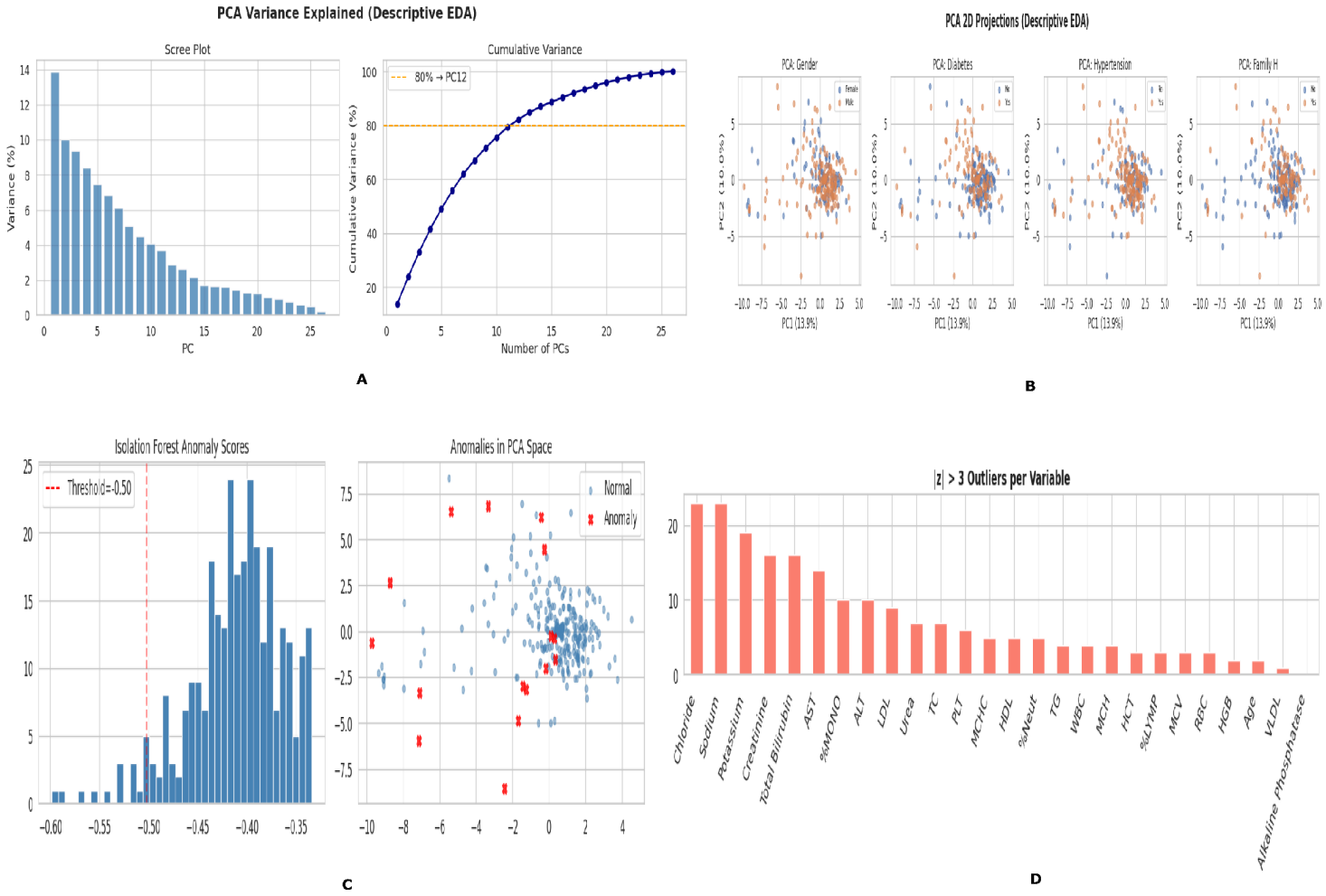


Figure 3

Dimensionality reduction and outlier/anomaly analysis

(A) Scree plot showing the percentage of variance explained by each principal component (left) and cumulative variance explained as a function of the number of components, with the threshold for 80% cumulative variance indicated (right). (B) Two-dimensional PCA projections (PC1 vs. PC2) of patients, colored by sex, diabetes status, hypertension status, and family history, respectively. (C) Distribution of Isolation Forest anomaly scores with the chosen classification threshold (left) and identified anomalies (red) plotted in PCA space relative to normal samples (blue) (right). (D) Number of univariate outliers (|z| > 3) detected for each variable.

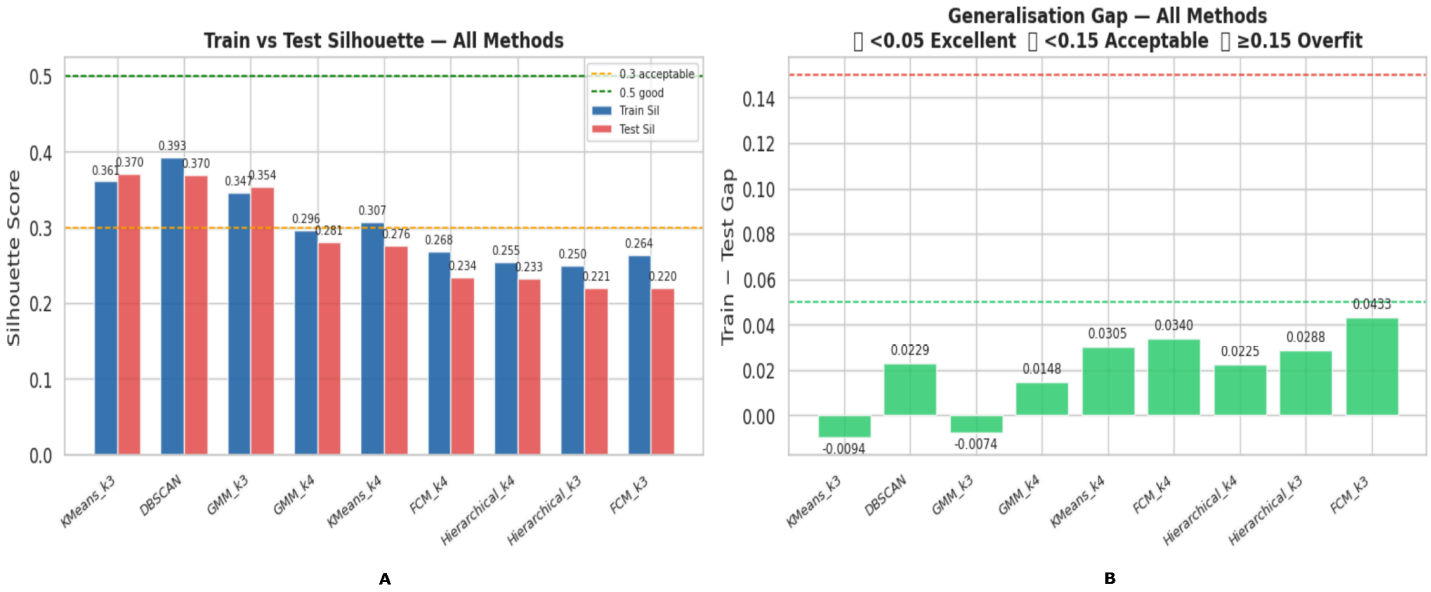


Figure 4

Clustering performance and generalization assessment across unsupervised methods
(A) Silhouette scores on the training and held-out test sets for each clustering method/configuration (KMeans, DBSCAN, GMM, Hierarchical, and Fuzzy C-Means, at k = 3 or k = 4), with reference lines denoting "acceptable" (0.3) and "good" (0.5) silhouette thresholds. (B) Generalization gap (train − test silhouette score) for each method, with reference lines denoting excellent (<0.05), acceptable (<0.15), and overfit (≥0.15) thresholds.

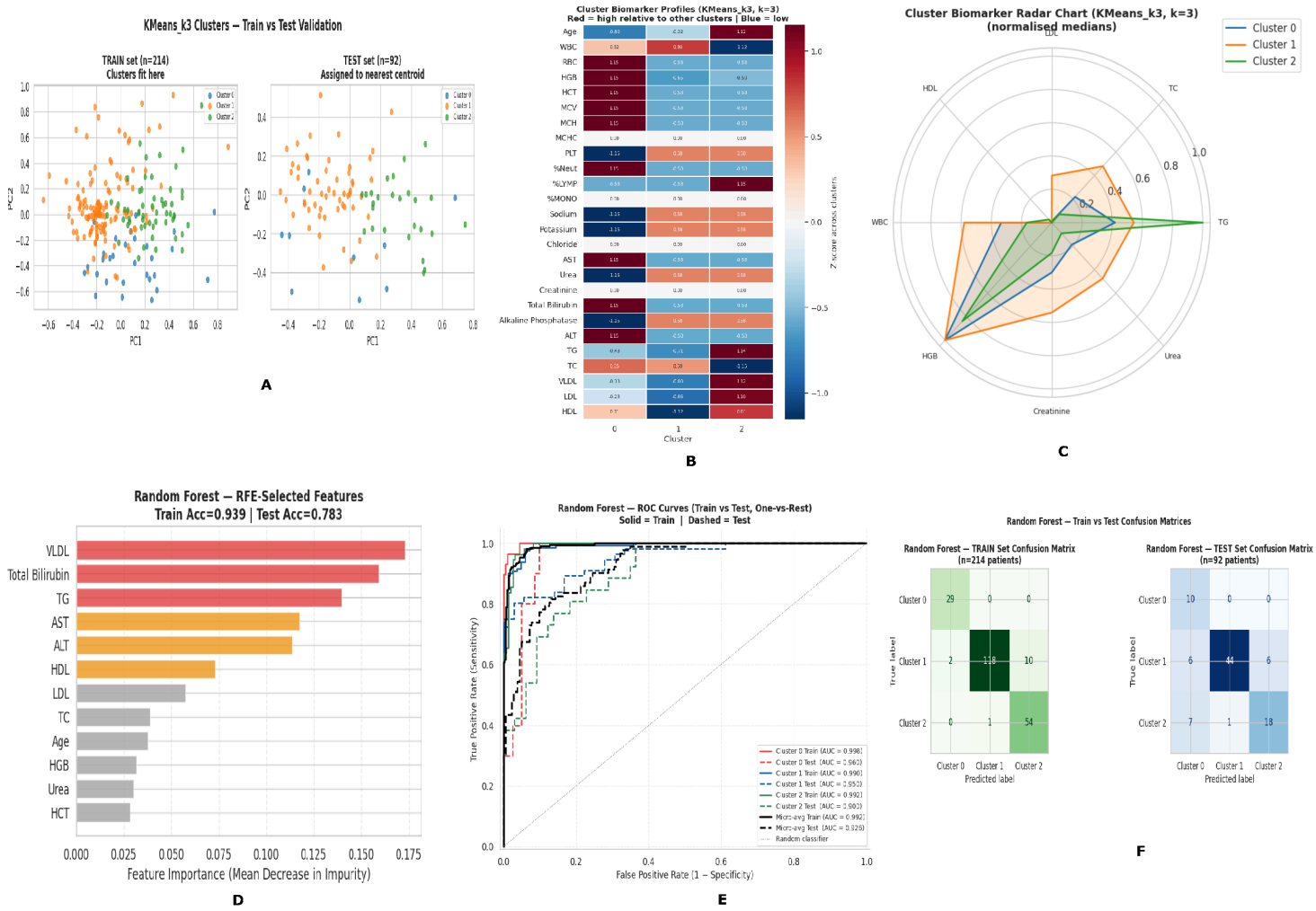


Figure 5

Random Forest classification of KMeans-derived patient clusters (k = 3)

(A) PCA scatter plots (PC1 vs. PC2) showing cluster assignments in the training set (clusters fit here) and test set (assigned to nearest centroid). (B) Heatmap of z-score–normalized biomarker profiles across the three clusters, highlighting relatively high (red) and low (blue) values per feature. (C) Radar chart of normalized median biomarker values illustrating the distinct multivariate profile of each cluster. (D) Feature importance (mean decrease in impurity) for RFE-selected predictors used in the Random Forest classifier; training and test accuracy are indicated above the panel. (E) One-vs-rest ROC curves with AUC values for each cluster, shown for the training and test sets. (F) Confusion matrices for the training set (n = 214) and test set (n = 92).

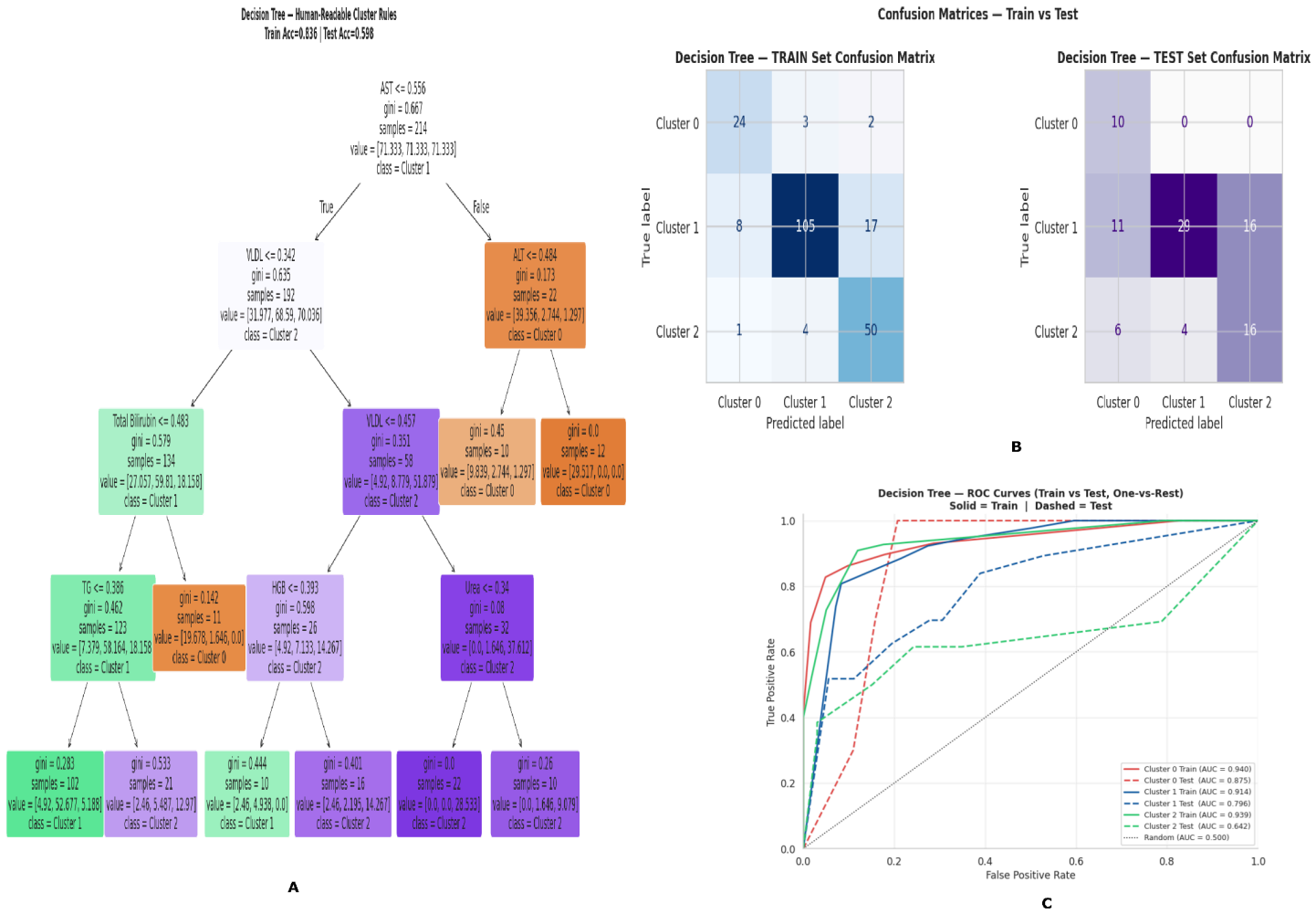


Figure 6

Decision tree model for interpretable cluster classification

(A) Decision tree structure showing the hierarchical splitting rules (based on AST, VLDL, ALT, Total Bilirubin, TG, HGB, and Urea) used to classify patients into clusters, with Gini impurity, sample counts, and class values at each node; training and test accuracies are shown above the panel. (B) Confusion matrices for the training and test sets. (C) One-vs-rest ROC curves with AUC values for each cluster on the training and test sets.



Figure 7

Fuzzy rule-based characterization of patient clusters
(A) Comparison of training and test set accuracy across the three modeling approaches (Random Forest, Decision Tree, and ExFuzzy). (B) Rule accuracy (left) and rule coverage, i.e., degree of support/percentage of patients covered (right), for each extracted fuzzy rule, grouped by cluster. (C) Sankey diagram mapping fuzzy rules to patient clusters, with flow width proportional to patient coverage and color indicating rule accuracy (green = high, orange = medium, red = low).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryDataS1.xlsx
- SupplementaryFigureS1.png
- SupplementaryInformationS1.docx
- SupplementaryTableS1.docx

- SupplementaryTableS2.docx