GALLSTONE DISEASE PREDICTION WITH EXPLAINABLE ARTIFICIAL INTELLIGENCE

ACOSTA, DELA CRUZ, PERILLO, RODRIGUEZ

CMSC 177

OVERVIEW

- Introduction
- Literary Review
- Methodology
- Result
- Conclusion
- Recommendation

BACKGROUND

- Gallstones are quite common and influenced by age, comorbidities, and body composition.
- The growth of clinical data creates opportunities for machine learning-based early prediction.
- Early and accurate diagnosis can greatly improve quality of life.

TRADITIONAL METHODS

- Diagnosis mostly through imaging after symptoms appear.
- Risk assessments are often subjective and limited.
- Blood tests alone aren't predictive.
- These approaches miss early or asymptomatic cases.

MACHINE LEARNING

- Offers early, data-driven prediction.
- Models used: Random Forest, SVM, Gradient Boosting, Logistic Regression
- Use of SHAP and LIME for model transparency and clinical trust.

EXPECTED CONTRIBUTION

- Goal: Predict gallstones using explainable ML
- Benefits: early detection, support decision-making.
- Steps: feature selection → model training → evaluation → interpretation

LITERATURE REVIEW

Machine Learning in Medical Diagnostics

Machine learning is being used more and more to accurately identify a variety of medical conditions, enabling quicker, data-driven diagnoses that enhance early intervention and lower errors.

Machine Learning for Gallstone Prediction

Gallstone risk can be reliably predicted by models that integrate metabolic data, inflammation markers, and body composition. They provide cheap, non-invasive substitutes for conventional imaging.

LITERATURE REVIEW

Explainable AI (XAI)

By clarifying the decision-making process of models, XAI tools such as SHAP and LIME develop clinician trust and promote the moral and open application of AI in healthcare.

XAI in Practice

Research on gallstones is still in early stages, but XAI has advanced knowledge of other illnesses. By using it in gallstone models, important risk factors are highlighted and individualized care is supported.

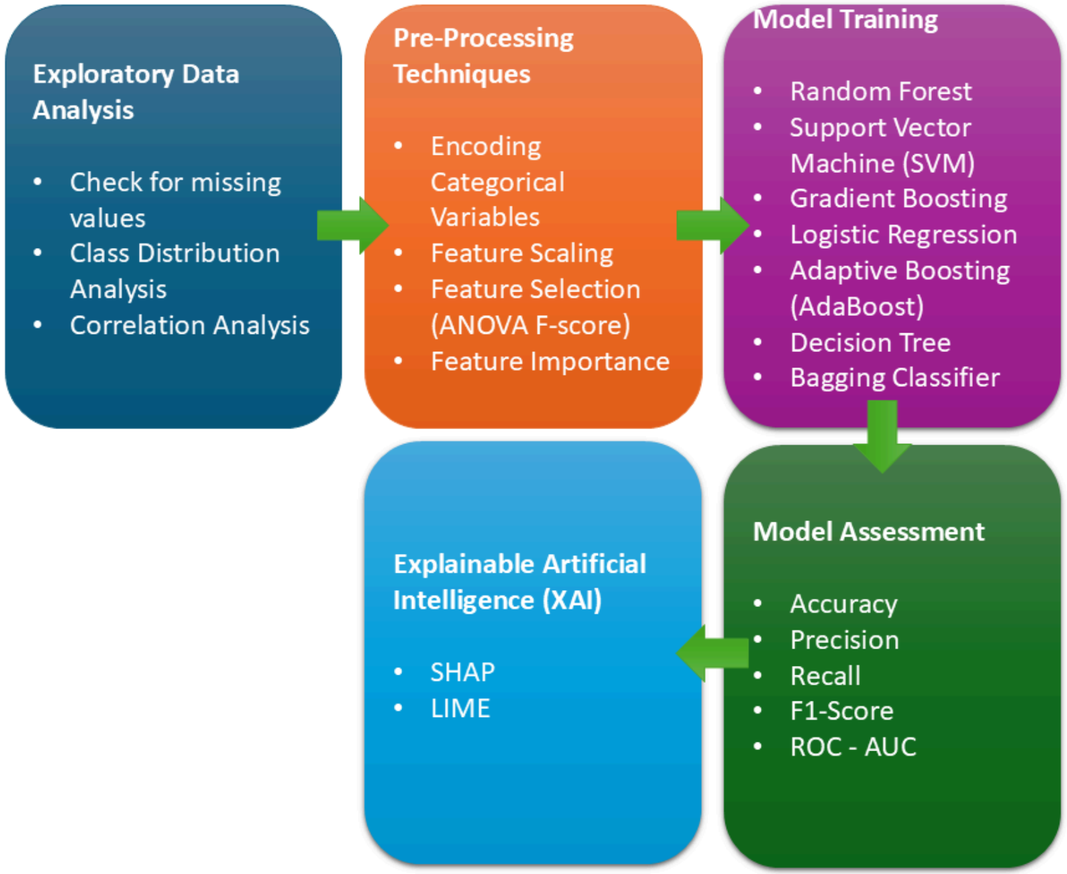
LITERATURE REVIEW

Effective Models and Features

Complex, unbalanced data can be effectively handled by algorithms such as Random Forest and Gradient Boosting. For improved performance and clarity, feature selection guarantees that models concentrate on the most crucial clinical variables.

METHODOLOGY

Machine Learning Pipeline



1. DATASET DESCRIPTION

The dataset used in this study was sourced from a publicly available repository for gallstone disease classification. It consists of 319 patient records, each containing real-valued clinical and laboratory features, with the target label Gallstone
Status indicating the presence or absence of gallstones.

The dataset was confirmed to have no missing values and was balanced across target classes. Each record contains features derived from medical parameters such as lean mass, extracellular water, and vitamin levels, which have shown predictive value in prior studies. This structure enabled robust machine learning development without the need for oversampling techniques.

2. DATA PRE-PROCESSING

Initial Checks & Findings

Checks Conducted:

- Missing values
- Duplicates
- Outlier presence
- Class distribution

Findings:

- No missing or duplicate values
- Dataset was balanced
- Clinical outliers were present but retained for relevance

Data Cleaning & Transformation

Cleaning Steps:

- Verified dataset integrity (319 complete records)
- No imputation or oversampling needed

Transformation:

- Applied standard scaling (zero mean, unit variance)
- Ensured uniformity across numeric clinical features

Notes on Outliers and Balance

- Outliers retained after boxplot analysis
- Dataset balance → no need for SMOTE
- Preprocessing supported model convergence and interpretability

2. DATA PRE-PROCESSING

Exploratory Data Analysis (EDA)

Descriptive Statistics

- Computed mean, median, standard deviation, min, and max
- Helped analyze key features like Lean Mass, ECF/TBW, and CRP
- Supported data normalization (standard scaling)

Visualization & Outlier Detection

- Used histograms and boxplots to explore distributions
- Assessed clinical outliers and retained them for relevance
- Helped validate class balance and variable spread

Correlation Analysis & Feature Selection

- Generated correlation heatmaps to explore inter-feature relationships
- Informed feature selection process using ANOVA F-value (SelectKBest)
- Threshold optimized via cross-validation for best model performance

Split Strategy

- Dataset divided into 70% training and 30% testing using stratified split
- Ensured balanced target distribution across training and test sets

Descriptive Statistics

```
print("\nDescriptive statistics for numerical features:")
# display.display(df[num_features].describe())
display(df[num_features].describe())
```

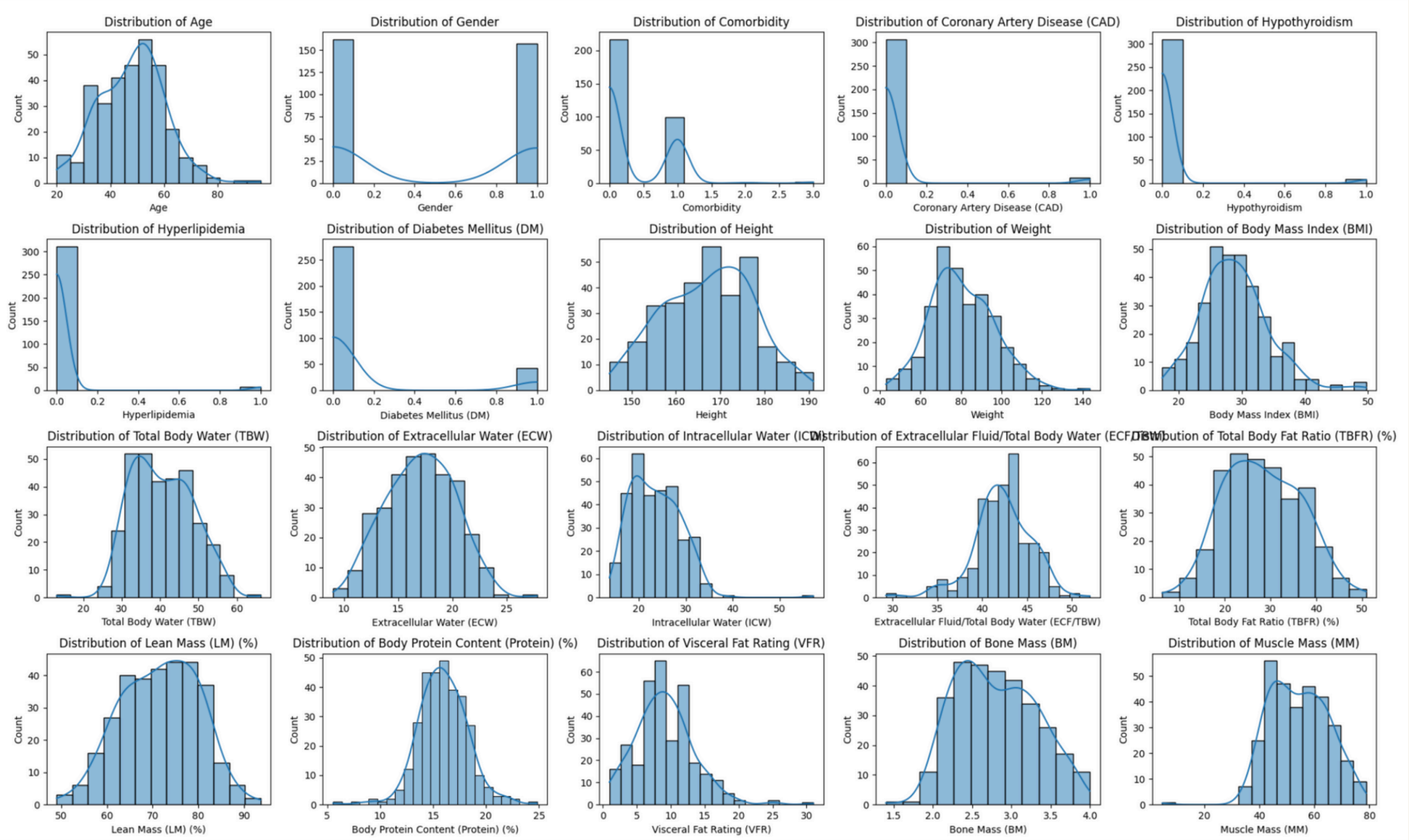
✓ 0.0s

Descriptive statistics for numerical features:

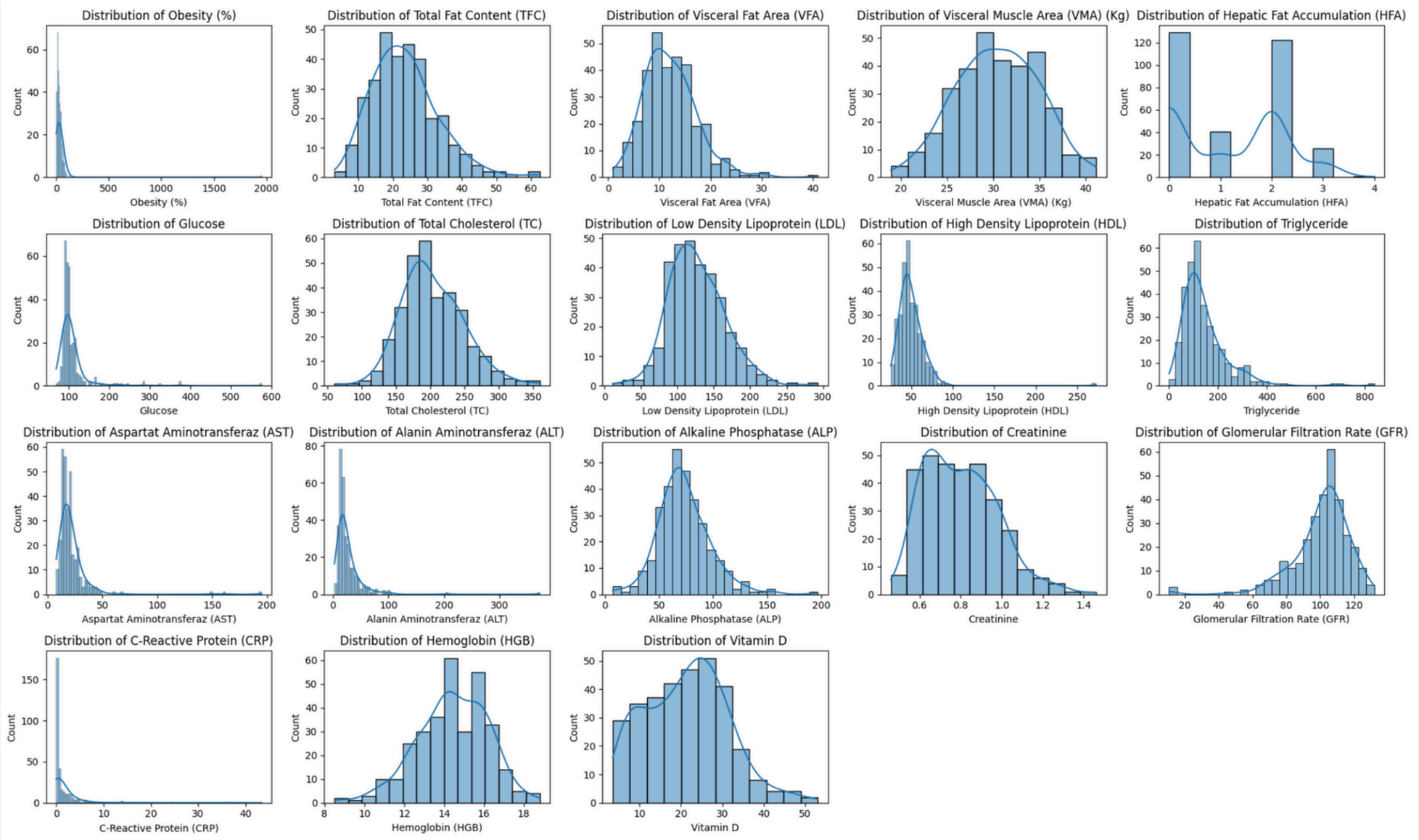
	Age	Gender	Comorbidity	Coronary Artery Disease (CAD)	Hypothyroidism	Hyperlipidemia	Diabetes Mellitus (DM)	Height	Weight	Body Mass Index (BMI)	...
count	319.000000	319.000000	319.000000	319.000000	319.000000	319.000000	319.000000	319.00000	319.000000	319.000000	...
mean	48.068966	0.492163	0.335423	0.037618	0.028213	0.025078	0.134796	167.15674	80.564890	28.877116	...
std	12.114558	0.500724	0.517340	0.190568	0.165841	0.156609	0.342042	10.05303	15.709069	5.313707	...
min	20.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	145.00000	42.900000	17.400000	...
25%	38.500000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	159.50000	69.600000	25.250000	...
50%	49.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	168.00000	78.800000	28.300000	...
75%	56.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	175.00000	91.250000	31.850000	...
max	96.000000	1.000000	3.000000	1.000000	1.000000	1.000000	1.000000	191.00000	143.500000	49.700000	...

High Density Lipoprotein (HDL)	Triglyceride	Aspartat Aminotransferaz (AST)	Alanin Aminotransferaz (ALT)	Alkaline Phosphatase (ALP)	Creatinine	Glomerular Filtration Rate (GFR)	C-Reactive Protein (CRP)	Hemoglobin (HGB)	Vitamin D
319.000000	319.000000	319.000000	319.000000	319.000000	319.000000	319.000000	319.000000	319.000000	319.000000
49.475549	144.502163	21.684953	26.855799	73.112539	0.800611	100.818903	1.853856	14.418182	21.401411
17.718701	97.904493	16.697605	27.884413	24.181069	0.176433	16.971396	4.989591	1.775815	9.981659
25.000000	1.390000	8.000000	3.000000	7.000000	0.460000	10.600000	0.000000	8.500000	3.500000
40.000000	83.000000	15.000000	14.250000	58.000000	0.650000	94.170000	0.000000	13.300000	13.250000
46.500000	119.000000	18.000000	19.000000	71.000000	0.790000	104.000000	0.215000	14.400000	22.000000
56.000000	172.000000	23.000000	30.000000	86.000000	0.920000	110.745000	1.615000	15.700000	28.060000
273.000000	838.000000	195.000000	372.000000	197.000000	1.460000	132.000000	43.400000	18.800000	53.100000

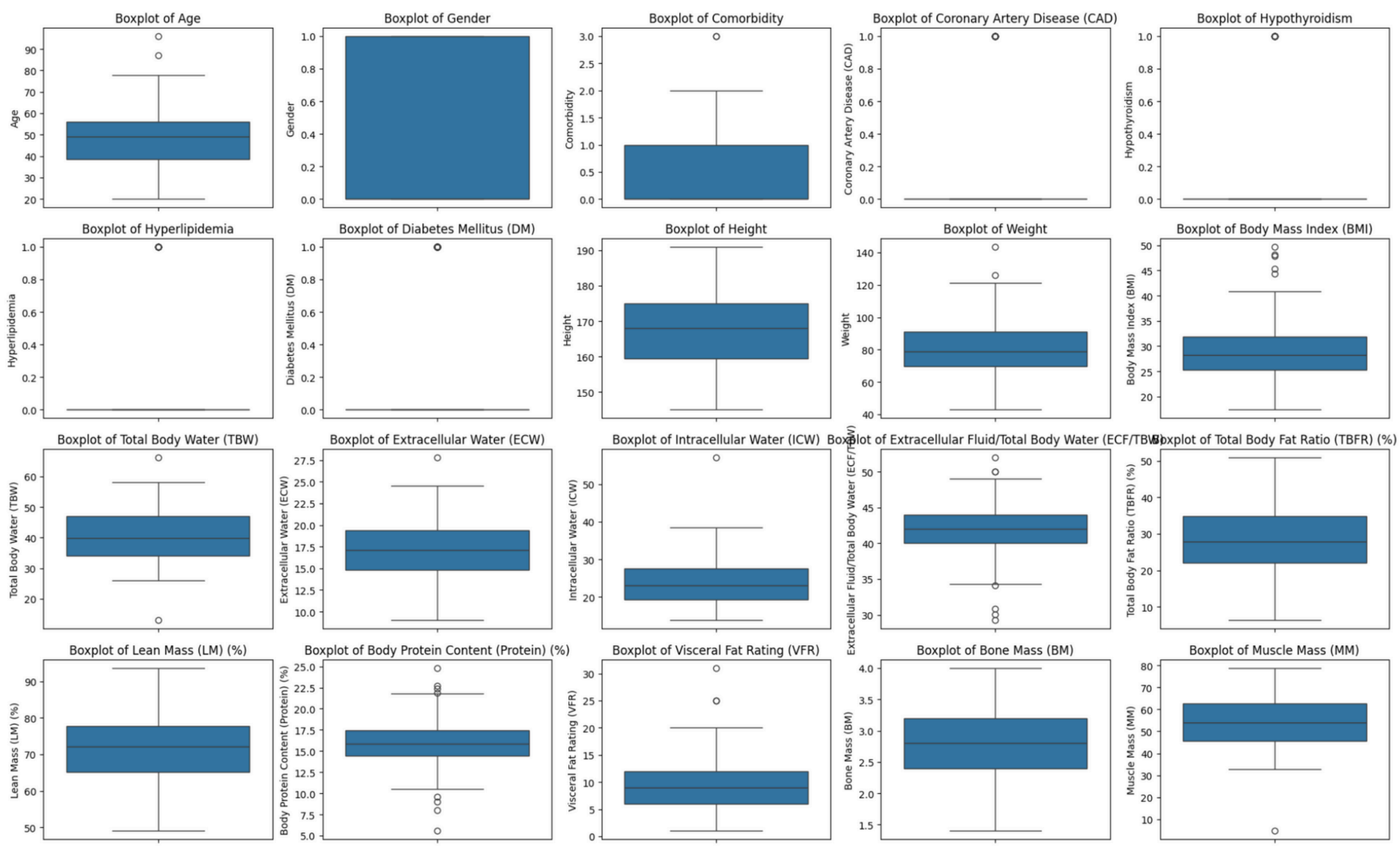
Histograms of Numerical Features



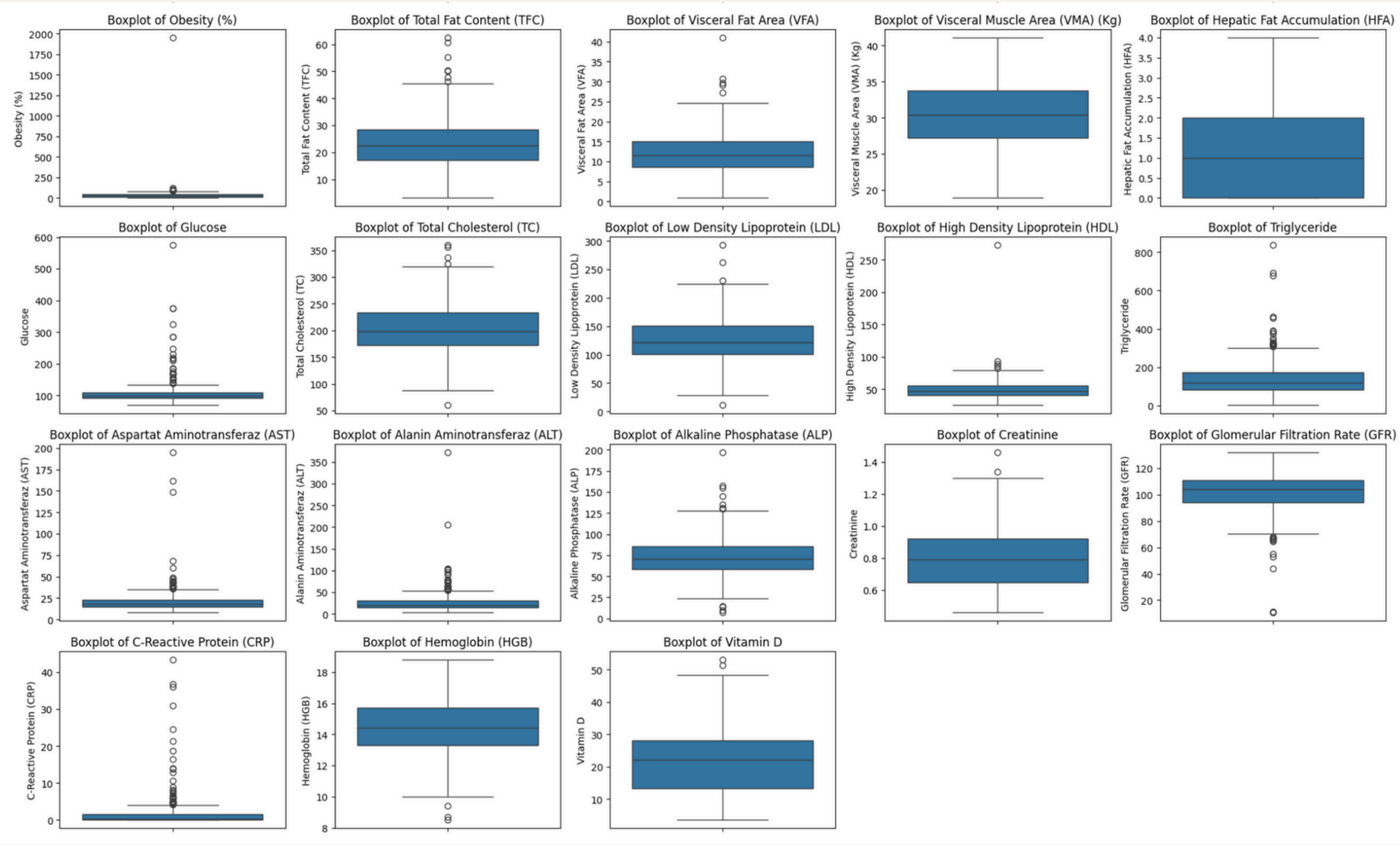
Histograms of Numerical Features

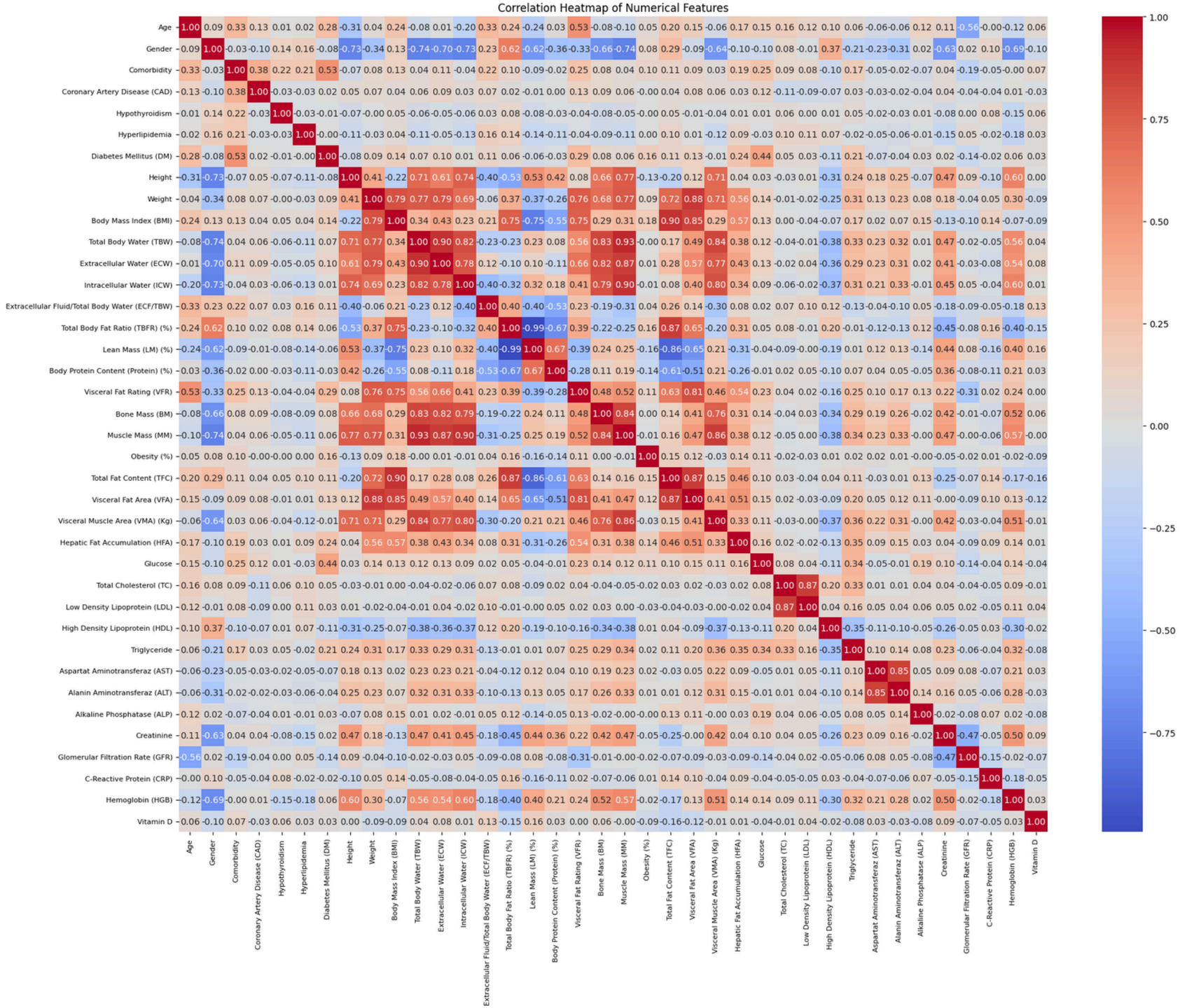


Boxplots of Numerical Features

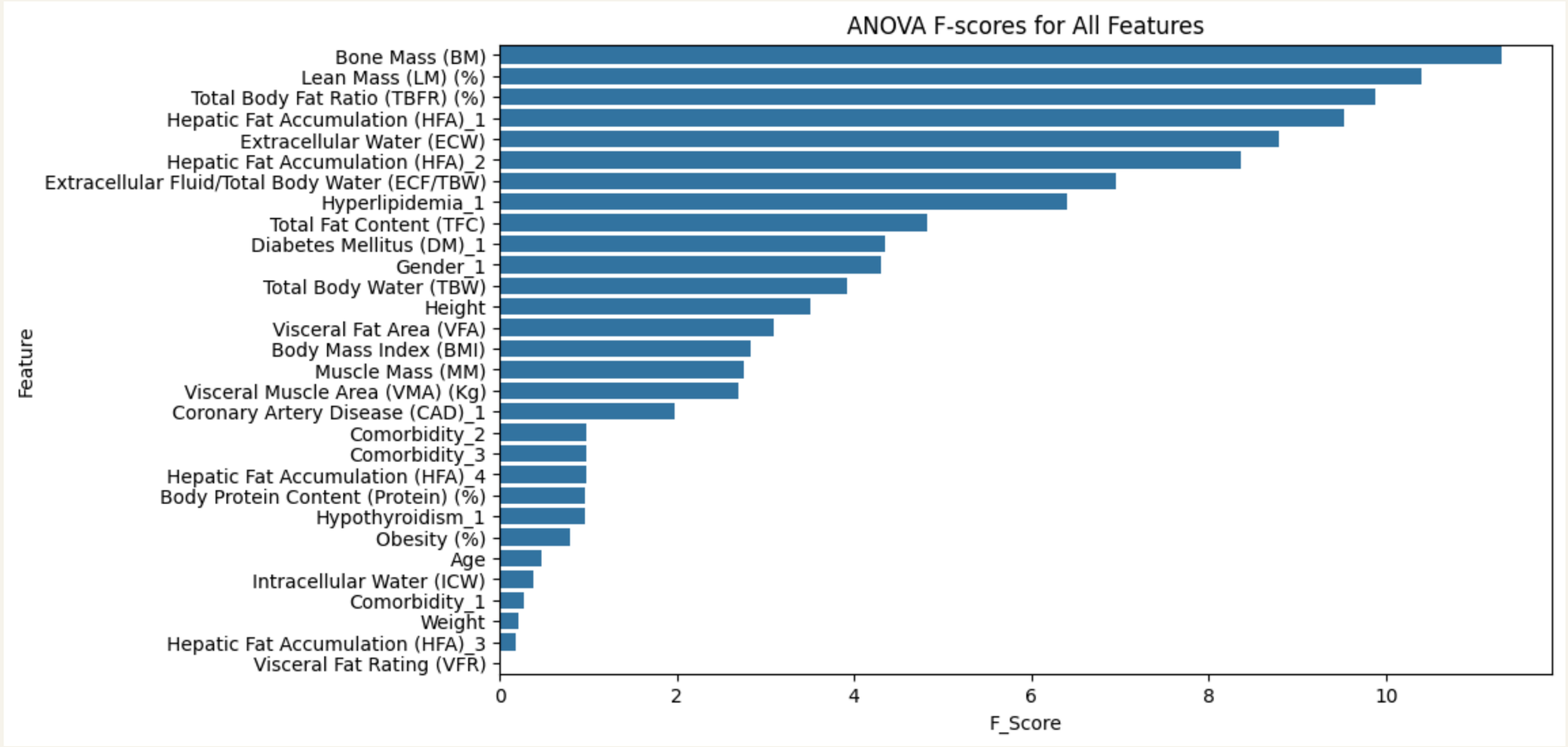


Boxplots of Numerical Features




Correlation Heatmap of Numerical Features
Age
Gender
Comorbidity
Coronary Artery Disease (CAD)
Hypothyroidism
Hyperlipidemia
Diabetes Mellitus (DM)
Height
Weight
Body Mass Index (BMI)
Total Body Water (TBW)
Extracellular Water (ECW)
Intracellular Water (ICW)
Extracellular Fluid/Total Body Water (ECF/TBW)
Total Body Fat Ratio (TBFR) (%)
Lean Mass (LM) (%)
Body Protein Content (Protein) (%)
Visceral Fat Rating (VFR)
Bone Mass (BM)
Muscle Mass (MM)
Obesity (%)
Total Fat Content (TFC)
Visceral Fat Area (VFA)
Visceral Muscle Area (VMA) (Kg)
Hepatic Fat Accumulation (HFA)
Glucose
Total Cholesterol (TC)
Low Density Lipoprotein (LDL)
High Density Lipoprotein (HDL)
Triglyceride
Aspartat Aminotransferaz (AST)
Alanin Aminotransferaz (ALT)
Alkaline Phosphatase (ALP)
Creatinine
Glomerular Filtration Rate (GFR)
C-Reactive Protein (CRP)
Hemoglobin (HGB)
Vitamin D
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

ANOVA F-Scores



ANOVA F-Score Threshold

- A threshold at the 75th percentile was set to retain the most relevant features.
- After application, the top 8 features were selected

MACHINE LEARNING MODELS

Train-Test Split

A Train-Test split of 70-30 was selected.

ML Algorithms

- Random Forest
- Support Vector Machine
- Gradient Boosting
- Logistic Regression
- Adaptive Boosting
- Decision Tree
- Bagging Classifier

Performance Metrics

- Accuracy
- Precision
- Recall
- FI-Score
- AUC

Validation

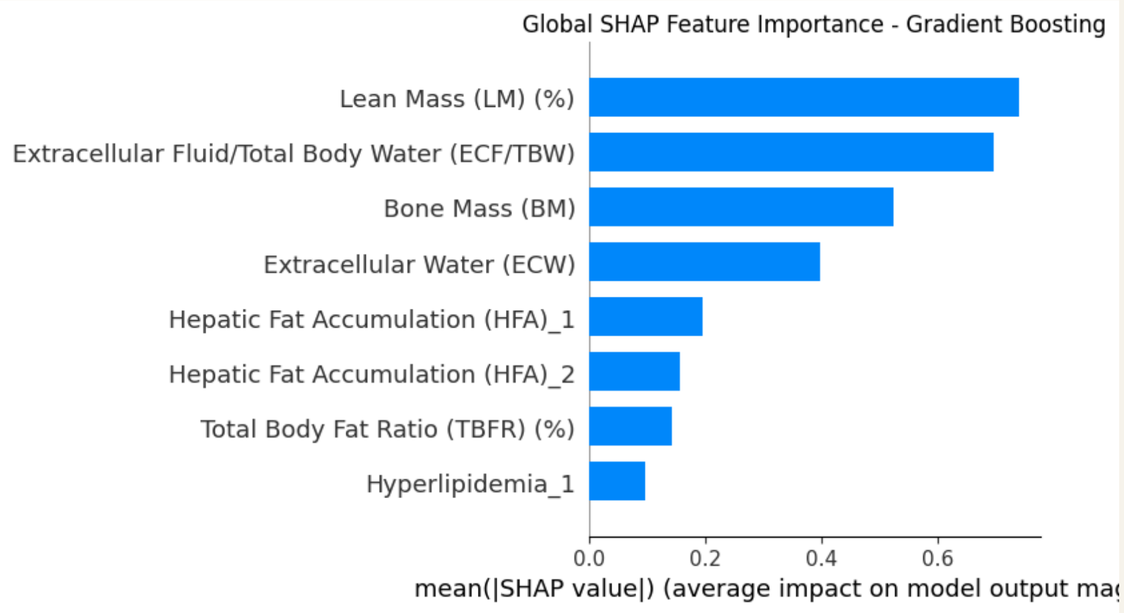
5-fold cross-validation was implemented to reduce overfitting

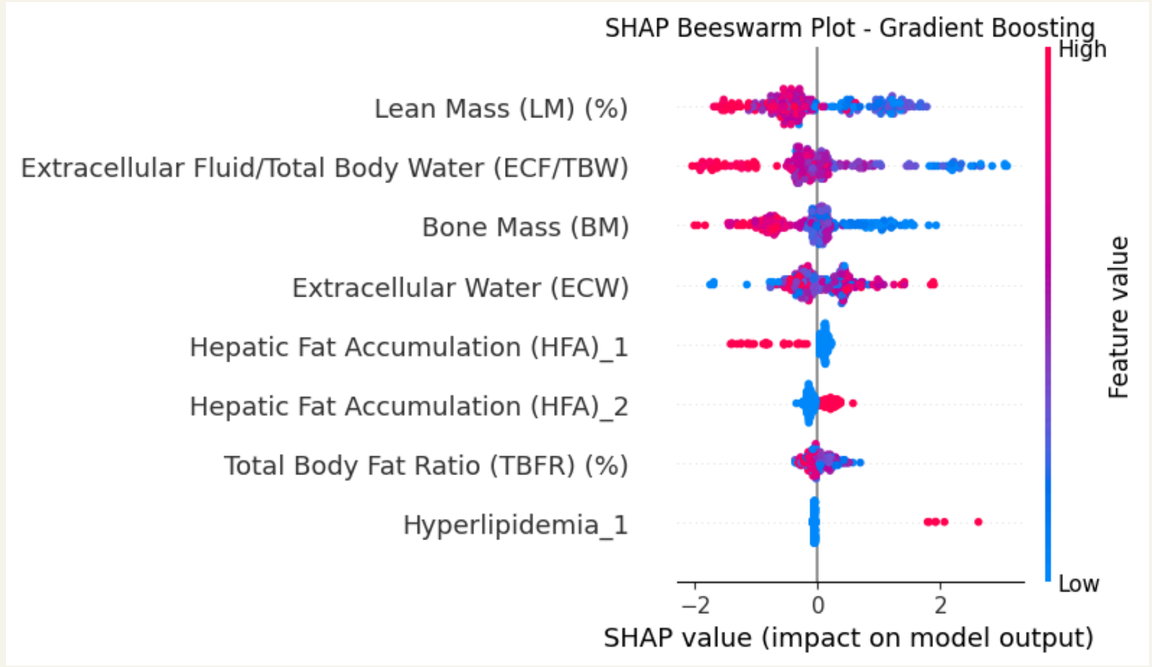
RESULT

ML Model	Accuracy	Precision	Recall	F1 Score	AUC
RF	0.75	0.79	0.69	0.73	0.79
SVM	0.75	0.94	0.53	0.68	0.82
GB	0.72	0.71	0.75	0.73	0.81
LR	0.77	0.81	0.69	0.75	0.86
AB	0.69	0.70	0.66	0.68	0.79
DT	0.73	0.74	0.72	0.73	0.73
BC	0.73	0.76	0.69	0.72	0.75

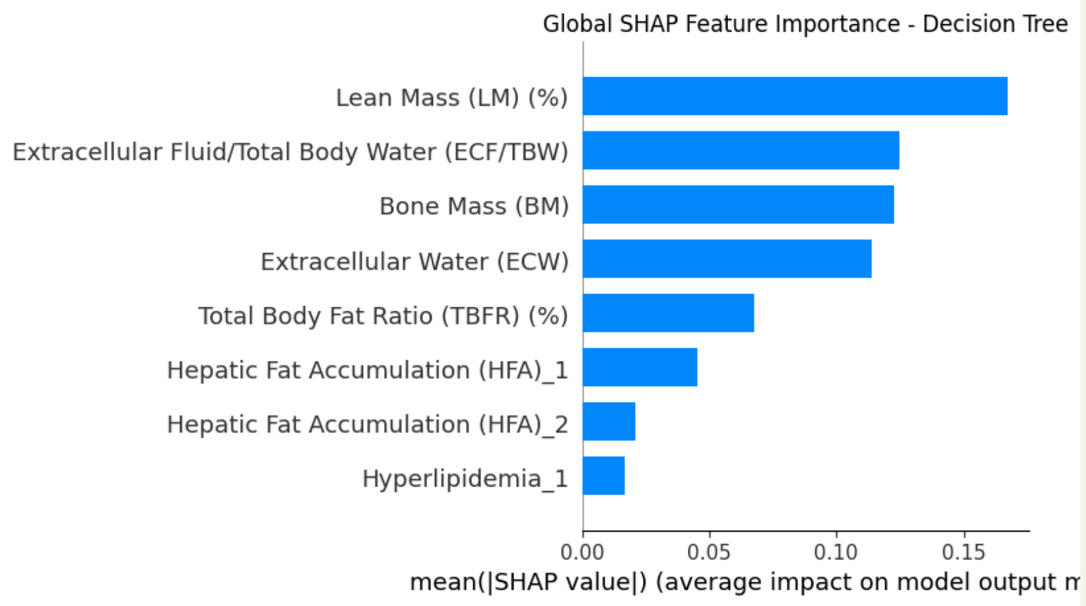
Best model in terms of relevant metrics (recall): Gradient Boosting

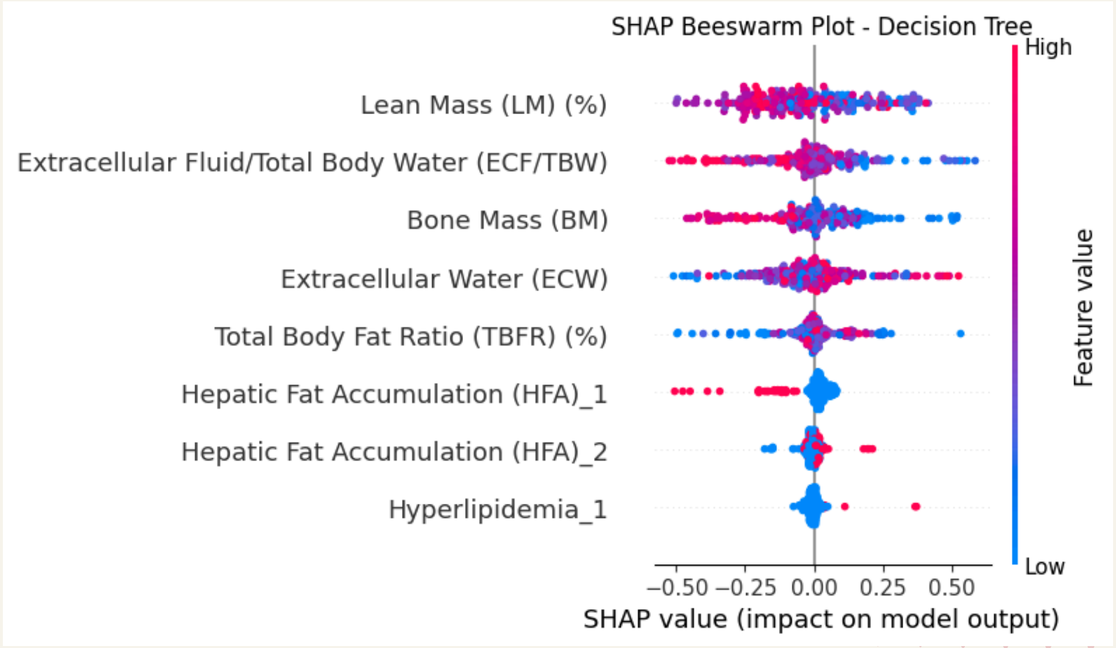
XAI EXPLAINERS: SHAP





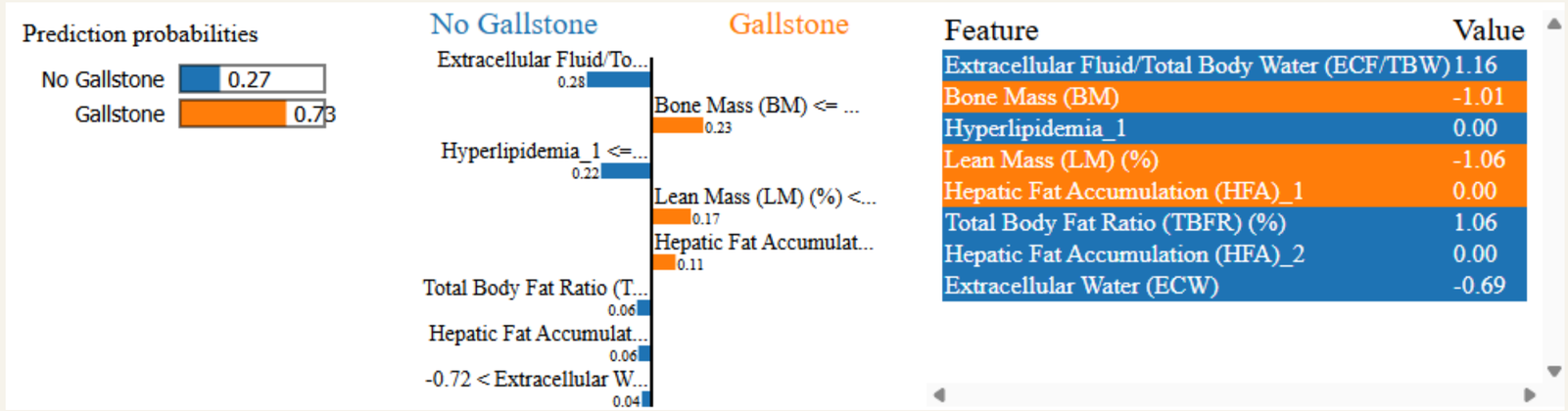
XAI EXPLAINERS: SHAP



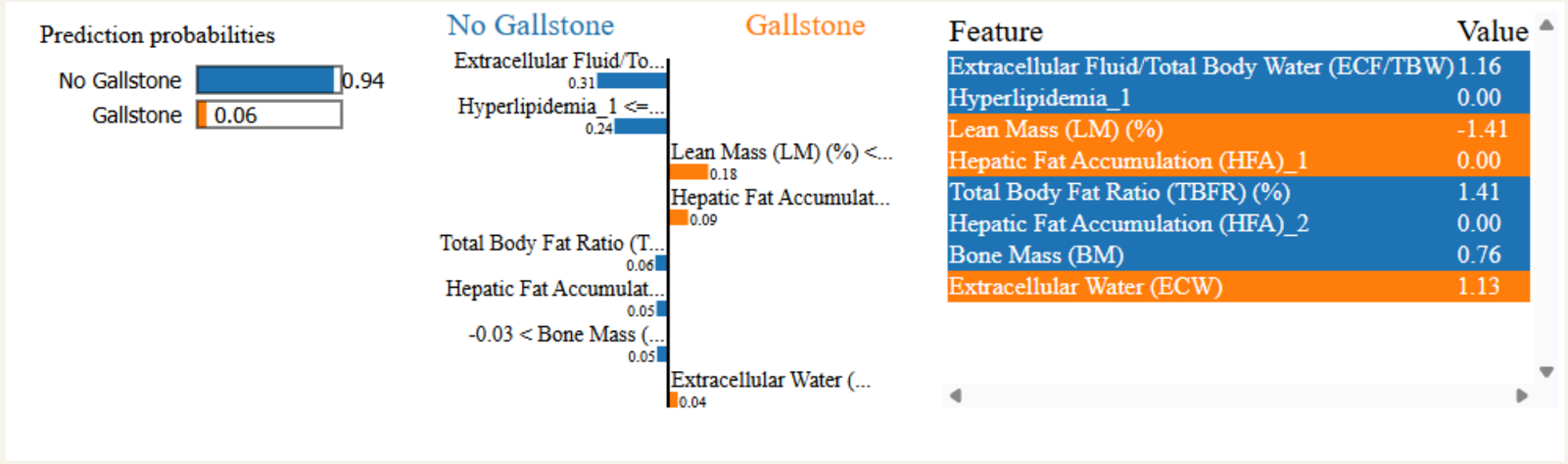


XAI EXPLAINERS: LIME

Positive Sample



Negative Sample



CONCLUSION

This study used clinical and physiological data to evaluate the predictive power of several machine learning models for gallstone disease. With a 75% recall rate, gradient boosting performed the best. By adding SHAP and LIME, the model became more transparent and the results were easier to interpret clinically. All things accounted for, the models hold potential for early diagnosis and individualized risk assessment in patient care.

RECOMMENDATION

Expand Dataset Diversity

To increase the generalizability and dependability of the model, incorporate data from more sources and larger populations.

Improve Preprocessing Techniques

Improve input quality and model performance by implementing more thorough data-cleaning and transformation techniques.

Incorporate Advanced Features

To find more profound predictive patterns, use imaging, genetic information, and longitudinal records.

Explore Additional Modeling Approaches

To possibly boost predictive power, try out new approaches and ensemble techniques like stacking.

Conduct External Validation

To confirm the models' applicability and practical impact, test them on independent datasets or in actual clinical settings.

REFERENCES

[1]X. Wang et al., "Global Epidemiology of Gallstones in the 21st Century: A Systematic Review and Meta-Analysis," Clinical Gastroenterology and Hepatology, vol. 22, no. 8, Feb. 2024, doi: https://doi.org/10.1016/j.cgh.2024.01.051.

[2]J. Zhang et al., "Association between metabolically healthy overweight/obesity and gallstones in Chinese adults," Association between metabolically healthy overweight/obesity and gallstones in Chinese adults, vol. 20, no. 1, Mar. 2023, doi: https://doi.org/10.1186/s12986-023-00741-4.

[3]GlobalRPH, "Obesity Classification And The Risk Of Gallstones," GlobalRPH, Nov. 14, 2023. https://globalrph.com/2023/11/obesity-classification-risk-of-gallstones

[4]İrfan Esen, H. Arslan, Selin Aktürk Esen, Mervenur Gülşen, Nimet Kültekin, and Oğuzhan Özdemir, "Early prediction of gallstone disease with a machine learning-based method from bioimpedance and laboratory data," Medicine, vol. 103, no. 8, pp. e37258–e37258, Feb. 2024, doi: https://doi.org/10.1097/md.0000000000037258.

[5]G. L.-H. Wong et al., "Novel machine learning models outperform risk scores in predicting hepatocellular carcinoma in patients with chronic viral hepatitis," JHEP Reports, vol. 4, no. 3, p. 100441, Mar. 2022, doi: https://doi.org/10.1016/j.jhepr.2022.100441.

[6]Md. A. Islam, Md. Z. H. Majumder, and Md. A. Hussein, "Chronic kidney disease prediction based on machine learning algorithms," Journal of Pathology Informatics, p. 100189, Jan. 2023, doi: https://doi.org/10.1016/j.jpi.2023.100189.

[7]Mădălina Maria Muraru, Zsuzsa Simó, and László Barna Iantovics, "Cervical Cancer Prediction Based on Imbalanced Data Using Machine Learning Algorithms with a Variety of Sampling Methods," Applied Sciences, vol. 14, no. 22, pp. 10085–10085, Nov. 2024, doi: https://doi.org/10.3390/app142210085.

[8]Razan Alkhanbouli, Hour, F. Alhosani, and M. Can, "The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions," BMC Medical Informatics and Decision Making, vol. 25, no. 1, Mar. 2025, doi: https://doi.org/10.1186/s12911-025-02944-6.

[9]V. Vimbi, N. Shaffi, and M. Mahmud, "Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection," Brain informatics, vol. 11, no. 1, Apr. 2024, doi: https://doi.org/10.1186/s40708-024-00222-1.

[10]S. Rao, S. Mehta, S. Kulkarni, H. Dalvi, N. Katre, and M. Narvekar, "A Study of LIME and SHAP Model Explainers for Autonomous Disease Predictions," IEEE Xplore, Dec. 01, 2022. https://ieeexplore.ieee.org/abstract/document/10037324?casa_token=KIrsa6-k2D4AAAAA:Rq4iHUKIVXsLXsceC-kiSC2VGkT4djYLbrsOZvbCQhxasKdGDlxIlE3V_9iMnI4KhQX34CmFgnQv

[11]J.-M. Yin, Y. Li, J.-T. Xue, G.-W. Zong, Z.-Z. Fang, and L. Zou, "Explainable machine learning-based prediction model for diabetic nephropathy," arXiv.org, 2023. https://arxiv.org/abs/2309.16730

[12]Y. A. Yarkın and A. Kalayci, "Gradient Boosting Decision Trees on Medical Diagnosis over Tabular Data," arXiv.org, 2024. https://arxiv.org/abs/2410.03705

[13]B. Ahmad, J. Chen, and H. Chen, "Feature selection strategies for optimized heart disease diagnosis using ML and DL models," arXiv.org, 2025. https://arxiv.org/abs/2503.16577

[14]A. S. Shaikh, R. M. Samant, K. S. Patil, N. R. Patil, and A. R. Mirkale, "Review on Explainable AI by using LIME and SHAP Models for Healthcare Domain," International Journal of Computer Applications, vol. 185, no. 45, pp. 18–23, 2023, Accessed: May 30, 2025. [Online]. Available: https://www.ijcaonline.org/archives/volume185/number45/32992-2023923263

[15]M. Panda and M. S. Ranjan, "Explainable artificial intelligence for Healthcare applications using Random Forest Classifier with LIME and SHAP," arXiv.org, 2023. https://arxiv.org/abs/2311.05665

UPM | 2025

THANK YOU