# A k-NN Based Analysis of Look-Alike Health Profiles to Identify Key Differences in Gallstone Formation

1st Ada Şevval Sari
*Computer Engineering*
220303023

2nd Hüseyin Kaya
*Computer Engineering*
220303036

3rd Mert Kiyar
*Computer Engineering*
220303038

*Abstract*—**Gallstone disease is one of the most prevalent gastrointestinal disorders worldwide and often progresses asymptomatically until severe complications occur. Current diagnostic approaches primarily rely on imaging techniques such as ultrasonography, which may limit early detection due to cost and accessibility constraints. In this study, we propose a machine learning-based, non-invasive decision support framework for predicting gallstone formation using routinely collected demographic, bioimpedance, and laboratory data. A clinically approved dataset consisting of 320 individuals and 38 features was obtained from Ankara VM Medical Park Hospital (2022–2023). Multiple supervised learning algorithms were evaluated, with a particular focus on the k-Nearest Neighbors (k-NN) classifier to analyze look-alike health profiles and identify critical differentiating factors. Model performance was assessed using accuracy, recall, specificity, F1-score, and ROC–AUC metrics. The results demonstrate that similarity-based learning approaches can effectively capture complex nonlinear relationships among clinical variables and offer competitive predictive performance. The proposed framework highlights the potential of data-driven risk assessment tools to support early diagnosis and preventive healthcare strategies for gallstone disease.**

**Keywords:** Gallstone Disease, Machine Learning, k-NN, Clinical Data Analysis, Decision Support Systems, Feature Importance

## I. Introduction

Gallstone disease (Fig. 1) is a prevalent hepatobiliary disorder resulting from supersaturation and crystallization of bile components, primarily cholesterol. The condition affects millions of individuals worldwide and represents a major cause of gastrointestinal morbidity. Global population studies indicate that gallstone prevalence ranges between 10% and 20%, with higher rates reported in Western regions [1]. Female sex, advanced age, and obesity have been consistently identified as strong risk factors for gallstone formation [2]. Metabolic alterations such as dyslipidemia, insulin resistance, and hepatic fat accumulation further contribute to biliary cholesterol imbalance, increasing the likelihood of stone formation [3]. Although many patients remain asymptomatic, untreated gallstones may progress to severe complications, including acute cholecystitis and biliary obstruction. As clinical manifestations are often delayed or nonspecific, early identification of at-risk individuals remains a critical challenge in preventive hepatobiliary care.

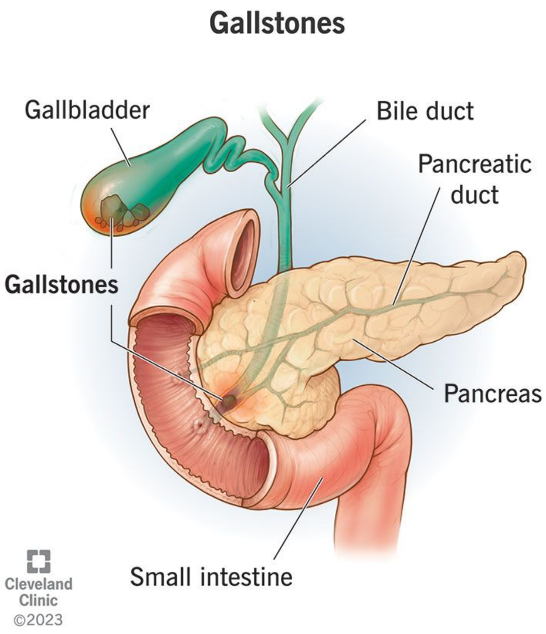

Fig. 1: Illustration of gallstone formation in the gallbladder.

Gallstone size (Fig. 2) plays a critical role in determining clinical severity, symptom development, and treatment strategies. Smaller stones (typically ¡5 mm) are more likely to migrate into the biliary ducts, increasing the risk of pancreatitis and choledocholithiasis [4]. In contrast, larger stones may remain asymptomatic for long periods but are associated with a higher probability of chronic inflammation and gallbladder dysfunction [5].

Stone growth rate is influenced by bile composition, cholesterol saturation, and gallbladder motility, with studies reporting measurable enlargement over time in untreated patients [6]. Understanding stone size dynamics is therefore essential for predicting clinical outcomes and guiding therapeutic decision-making.

Although ultrasonography remains the diagnostic gold standard, it may not be suitable for large-scale screening [7]. Recent advances in artificial intelligence and clinical data mod-

Fig. 2: Macroscopic view of extracted gallstones.

eling offer potential for non-invasive screening tools capable of identifying early risk determinants directly from patient features.

Despite progress, few studies have attempted to predict gallstone formation using machine learning models trained on routine clinical measurements. This motivates our investigation into the effectiveness of k-NN and other supervised classifiers for gallstone prediction.

## II. Related Work

A substantial number of epidemiological studies have examined metabolic and demographic risk factors for gallstone disease. Early research identified strong associations with obesity, triglyceride levels, and cardiovascular risk markers [3], [8]. More recent studies investigated hepatic fat accumulation and metabolic syndrome as contributing factors [9], [10].

Dietary influences have also been examined. Research indicates risk modulation through legume intake [11], high carbohydrate diets [12], and variations in lipid metabolism [13].

Genetic and population-based studies have shown significant variation across ethnic groups [14], [15]. Clinical studies have additionally linked gallstone formation to altered hepatic metabolism and biliary secretion processes [16].

Machine learning has seen increasing use in medical classification problems, including hepatobiliary disease, yet its application to gallstone prediction remains limited. Motivated by this gap, our work evaluates k-NN similarity modeling alongside multiple machine learning baselines.

## III. Dataset Description

The dataset employed in this study was retrospectively collected from Ankara VM Medical Park Hospital and includes clinical records of 320 individuals, 161 of whom were diagnosed with gallstone disease, examined between 2022 and 2023. Ethical approval was obtained prior to data acquisition, and all data were anonymized to ensure patient confidentiality. The dataset was specifically curated to investigate gallstone formation using non-invasive, routinely collected measurements, making it well-suited for machine learning-based risk prediction.

Each patient record consists of 38 independent variables spanning demographic characteristics, comorbidity indicators, bioimpedance-derived body composition metrics, and biochemical laboratory test results. The dependent variable represents gallstone status, indicating whether gallstones were detected during clinical evaluation.

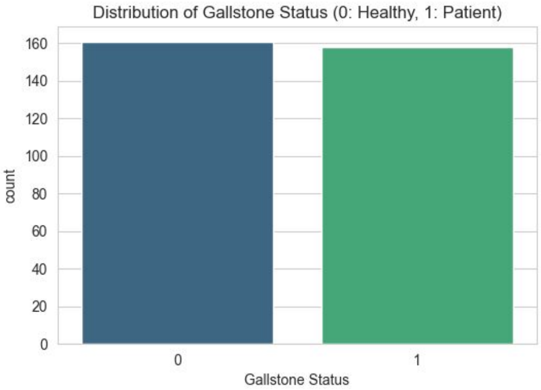

Fig. 3: Data distribution graph

### A. Target Variable

The target variable of the dataset is *Gallstone Status*, a binary indicator reflecting the presence or absence of gallstones. It is encoded as follows:

- 0: Gallstones present
- 1: Gallstones absent

This binary formulation enables the application of supervised classification algorithms and supports clinically interpretable outcome prediction.

### B. Demographic Variables

Demographic attributes provide essential information about baseline patient characteristics that are known to influence gallstone risk. In this study, five demographic variables were included: age, gender, height, weight, and Body Mass Index (BMI). Age and gender are represented as integer and categorical values, respectively, and capture well-established epidemiological trends such as increased gallstone prevalence among older adults and females. Height and weight are included as continuous anthropometric measurements, while BMI, derived from weight-to-height ratio, serves as a widely recognized indicator of adiposity and metabolic status. Collectively, these variables enable the identification of age-related and sex-related metabolic differences that may contribute to gallstone formation.

### C. Comorbidity and Disease Indicators

The dataset incorporates multiple disease-related variables that are clinically associated with biliary dysfunction and systemic metabolic burden. These include **Comorbidity count**, **Coronary Artery Disease (CAD)**, **Hypothyroidism**, **Hyperlipidemia**, and **Diabetes Mellitus (DM)**. Comorbidity is represented as an ordinal measure indicating the number of co-existing chronic conditions, while the remaining variables are encoded as binary indicators reflecting the presence or absence of each condition. Together, these features capture underlying inflammatory, endocrine, and cardiovascular influences that may predispose individuals to gallstone development. Their inclusion provides an important clinical context for examining the indirect metabolic pathways involved in bile component imbalance and gallbladder motility.

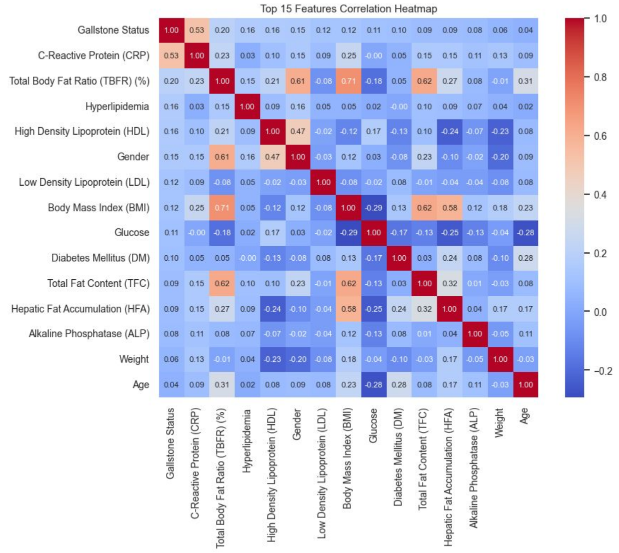

Fig. 4: Top 15 Features Correlation Heatmap

### D. Bioimpedance and Body Composition Features

A substantial proportion of the dataset consists of quantitative body composition measurements obtained through bioimpedance analysis. These features provide detailed insight into fluid distribution and adipose–muscle balance, including **Total Body Water (TBW)**, **Extracellular Water (ECW)**, **Intracellular Water (ICW)**, and the **ECF/TBW ratio**. Additionally, variables such as **Total Body Fat Ratio (TBFR)**, **Total Fat Content (TFC)**, **Lean Mass (LM)**, **Muscle Mass (MM)**, and **Visceral Muscle Area (VMA)** reflect structural and metabolic composition. Indicators of intra-abdominal adiposity, particularly **Visceral Fat Rating (VFR)** and **Visceral Fat Area (VFA)**, are especially relevant given the strong clinical association between visceral fat accumulation and biliary cholesterol supersaturation. The dataset further includes **Bone Mass (BM)**, **Body Protein Content**, and **Obesity percentage**, offering a comprehensive anthropometric profile that complements metabolic risk evaluation.

### E. Hepatic Fat Accumulation

Hepatic steatosis severity is represented by the **Hepatic Fat Accumulation (HFA)** variable, encoded as an ordinal scale from 0 (no fat accumulation) to 4 (very severe). This feature allows the model to assess the relationship between liver fat content and gallstone formation, reflecting evidence that impaired hepatic lipid metabolism contributes to cholesterol-enriched bile synthesis and gallbladder dysmotility.

### F. Laboratory Measurements

The dataset also includes routine biochemical markers that provide metabolic, hepatic, renal, and inflammatory information. These comprise **Glucose**, **Triglyceride**, **Total Cholesterol (TC)**, **LDL**, and **HDL**, which characterize lipid and glucose metabolism. Liver function is described by enzyme levels including **AST**, **ALT**, and **ALP**, while renal status is represented by **Creatinine** and **Glomerular Filtration Rate (GFR)**. Systemic inflammation is monitored using **C-Reactive Protein (CRP)**, and overall physiological status is supplemented with **Hemoglobin (HGB)** and **Vitamin D** levels. Collectively, these laboratory values enhance the model's ability to capture subtle biochemical differences associated with gallstone risk.

All variables in the dataset are complete, meaning no missing values were identified, thereby eliminating the need for imputation and ensuring consistent analytical integrity across all measurements.

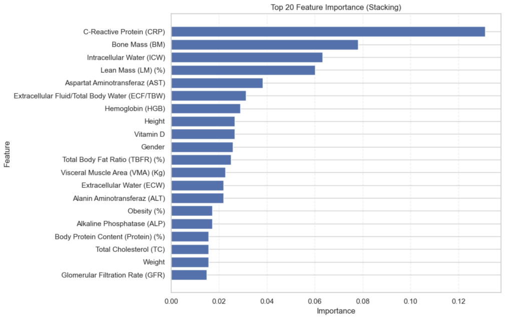

Fig. 5: Top 20 Feature Importance (Stacking)

## IV. Data Preprocessing

Prior to model training, a comprehensive data preprocessing pipeline was applied to ensure data consistency, numerical stability, and optimal model performance. Since the dataset contains heterogeneous variables originating from demographic, bioimpedance, and laboratory measurements, preprocessing plays a critical role, particularly for distance-based algorithms such as k-Nearest Neighbors.

### A. Categorical Variable Encoding

Categorical variables were transformed into numerical representations using label encoding. This approach assigns an integer value to each category while preserving ordinal relationships where applicable.

The following encoding schemes were applied:

- **Gender**: 0 (Male), 1 (Female)
- **Comorbidity**: 0 (None), 1 (One), 2 (Two), 3 (Three or more)
- **Binary disease indicators** (CAD, Hypothyroidism, Hyperlipidemia, DM): 0 (No), 1 (Yes)
- **Hepatic Fat Accumulation (HFA)**: Ordinal scale from 0 (no fat accumulation) to 4 (very severe)

Label encoding was preferred over one-hot encoding to avoid unnecessary dimensionality expansion and to maintain ordinal structure in clinically graded variables.

### *B. Feature Scaling and Normalization*

All numerical features were normalized to a common scale using Min–Max normalization. This step is particularly important for distance-based classifiers, as features measured on larger numerical ranges can dominate similarity calculations.

Each feature $x$ was scaled to the interval $[0, 1]$ using the following transformation:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

where $x_{\min}$ and $x_{\max}$ denote the minimum and maximum values of the feature within the dataset, respectively.

This normalization ensures equal contribution of all features during distance computation and improves convergence behavior for gradient-based learning models.

### *C. Handling of Missing Values*

The dataset contains no missing values across all variables. Therefore, no imputation techniques were required. The absence of missing data eliminates potential bias introduced by estimation-based filling methods and ensures data integrity throughout the modeling process.

### *D. Dataset Splitting*

To evaluate the generalization capability of the models, the dataset was partitioned into training and test subsets. The training set was used for model fitting and hyperparameter tuning, while the test set was reserved for unbiased performance evaluation.

### *E. Cross-Validation Strategy*

To mitigate overfitting and enhance robustness, k-fold cross-validation was employed during model training. The dataset was divided into $k$ disjoint subsets, and each model was trained $k$ times using $k-1$ folds for training and one fold for validation.

The average performance across all folds was used as the final validation score:

$$\text{CV}_{score} = \frac{1}{k} \sum_{i=1}^{k} \text{Performance}_i \tag{2}$$

This strategy provides a more reliable estimate of model performance, particularly in clinical datasets where generalizability is critical.

### *F. Impact on Distance-Based Learning*

Since k-NN relies directly on distance computations in feature space, preprocessing steps such as normalization and encoding are essential. Without proper scaling, variables such as visceral fat area or laboratory values could disproportionately influence distance calculations. The applied preprocessing pipeline ensures that similarity comparisons between patient profiles are both mathematically sound and clinically meaningful.

## V. Methodology

This study investigates the effectiveness of multiple supervised machine learning algorithms for predicting gallstone formation using clinical data. The selected methods include both classical and advanced classifiers to ensure a comprehensive comparative analysis. Each algorithm was chosen based on its theoretical strengths, interpretability, and suitability for structured medical datasets.

### *A. k-Nearest Neighbors (k-NN)*

The k-Nearest Neighbors algorithm is a non-parametric, instance-based learning method that classifies a query sample based on the majority class of its $k$ closest neighbors in the feature space. Similarity between instances was measured using the Euclidean distance:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{n} (x_j - x_{ij})^2} \tag{3}$$

where $x$ represents the query instance, $x_i$ denotes a training sample, and $n$ is the number of features. The value of $k$ was optimized through cross-validation to balance bias and variance. Due to its similarity-based nature, k-NN is particularly effective in identifying look-alike patient profiles.

### *B. Multilayer Perceptron (MLP)*

The Multilayer Perceptron is a feed-forward artificial neural network capable of modeling complex nonlinear relationships. Each neuron computes a weighted sum of its inputs followed by a nonlinear activation function:

$$h_j = f\left(\sum_{i=1}^{n} w_{ij} x_i + b_j\right) \tag{4}$$

The final output layer applies a sigmoid activation to estimate the probability of gallstone presence:

$$\hat{y} = \sigma\left(\sum_{j=1}^{m} v_j h_j + c\right) \tag{5}$$

where $w_{ij}$ and $v_j$ denote network weights, and $b_j$ and $c$ represent bias terms.

### *C. Support Vector Machines (SVM)*

Support Vector Machines aim to find an optimal hyperplane that maximally separates classes in a high-dimensional feature space. For linearly separable data, the decision boundary is defined as:

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \tag{6}$$

To handle nonlinear patterns, kernel functions were employed to project data into higher-dimensional spaces, enabling effective separation of complex clinical patterns.

### D. Decision Tree

Decision Trees classify data by recursively partitioning the feature space based on impurity measures. The Gini impurity index was used to determine optimal splits:

$$Gini = 1 - \sum_{i=1}^{C} p_i^2 \tag{7}$$

where $p_i$ is the proportion of samples belonging to class $i$ at a given node. Decision Trees provide intuitive interpretability by revealing feature-based decision paths.

### E. Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees using bootstrap sampling and random feature selection. The final prediction is obtained through majority voting:

$$\hat{y} = \text{mode}\left(h_1(x), h_2(x), \ldots, h_n(x)\right) \tag{8}$$

This approach improves generalization performance and reduces overfitting compared to single-tree models.

### F. Adaptive Boosting (AdaBoost)

AdaBoost combines multiple weak classifiers into a strong ensemble by iteratively reweighting misclassified samples. The weight assigned to each classifier is computed as:

$$\alpha_t = \frac{1}{2} \ln \left( \frac{1 - \epsilon_t}{\epsilon_t} \right) \tag{9}$$

where $\epsilon_t$ denotes the classification error of the $t$-th weak learner. AdaBoost enhances sensitivity to hard-to-classify instances.

### G. Gradient Boosting

Gradient Boosting builds models sequentially by fitting new learners to the residual errors of previous models. The model update rule is expressed as:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{10}$$

where $\gamma_m$ represents the learning rate and $h_m(x)$ denotes the newly added weak learner.

### H. Naïve Bayes

Naïve Bayes is a probabilistic classifier based on Bayes' theorem and assumes conditional independence among features:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \tag{11}$$

Despite its simplicity, Naïve Bayes often performs competitively in medical classification tasks with well-structured data.

### I. Logistic Regression

Logistic Regression models the probability of class membership using a logistic function:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\mathbf{w} \cdot \mathbf{x} + b)}} \tag{12}$$

This method serves as a baseline classifier and provides interpretable coefficients representing the influence of each feature.

### J. Model Selection Strategy

All models were trained and evaluated under identical preprocessing and validation conditions. Performance comparisons were conducted using standardized evaluation metrics, and the best-performing model was selected based on overall predictive accuracy and clinical interpretability.

## VI. Flowcharts And Pseudo-Code of the Proposed Framework

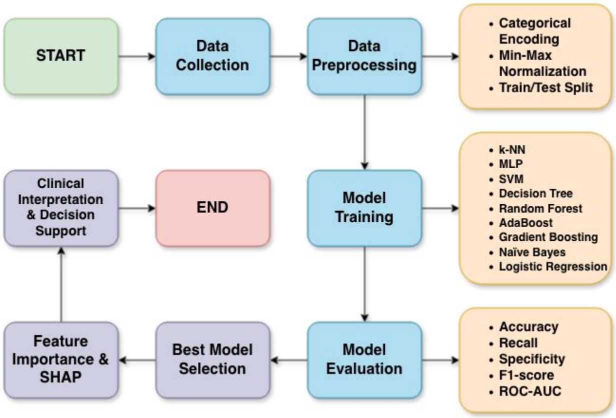

Fig. 6: A flowchart about this work

```
Input: Clinical dataset D
Output: Trained prediction model and
performance metrics

1. Load dataset D
2. Encode categorical features
3. Normalize numerical features
4. Split D into training and test sets
5. For each model M in model list:
 a. Train M on training set
 b. Validate M using cross-validation
 c. Evaluate M on test set
6. Compare models using performance metrics
7. Select best-performing model
```

## VII. Evaluation Metrics

To comprehensively evaluate model performance in a clinical classification setting, multiple evaluation metrics were employed. These metrics provide complementary perspectives on predictive accuracy, sensitivity to gallstone presence, and overall diagnostic reliability.

### A. Accuracy

Accuracy measures the proportion of correctly classified instances among all samples:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

where $TP$ denotes true positives, $TN$ true negatives, $FP$ false positives, and $FN$ false negatives.

### B. Recall (Sensitivity)

Recall, also referred to as sensitivity, measures the ability of the model to correctly identify patients with gallstones:

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

High recall is particularly important in medical screening tasks to minimize missed diagnoses.

### C. Specificity

Specificity evaluates the model's ability to correctly identify individuals without gallstones:

$$Specificity = \frac{TN}{TN + FP} \tag{15}$$

This metric is critical for reducing false-positive diagnoses and unnecessary clinical interventions.

### D. F1-Score

The F1-score represents the harmonic mean of precision and recall, balancing false positives and false negatives:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{16}$$

where precision is defined as:

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

### E. Receiver Operating Characteristic (ROC) and AUC

The ROC curve illustrates the trade-off between true positive rate and false positive rate across varying decision thresholds. The Area Under the Curve (AUC) provides a threshold-independent measure of classification performance, where values closer to 1 indicate superior discriminative ability.

### F. Feature Importance and Model Interpretability

Understanding the contribution of individual features is essential for clinical interpretability and trust in machine learning-based decision support systems. To achieve this, feature importance analysis was conducted using both model-specific and model-agnostic approaches.

For tree-based ensemble models, feature importance was derived based on impurity reduction, reflecting how frequently and effectively a feature contributes to decision-making across the ensemble.

Additionally, SHapley Additive exPlanations (SHAP) were employed to provide a unified and theoretically grounded interpretation framework. SHAP values quantify the marginal contribution of each feature to the model output by computing Shapley values from cooperative game theory.

Formally, the SHAP value for feature $i$ is defined as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[ f(S \cup \{i\}) - f(S) \right] \tag{18}$$

where $F$ denotes the set of all features and $f(S)$ represents the model prediction using feature subset $S$.

This approach enables both global and local interpretability by identifying which clinical variables most strongly influence gallstone risk predictions across the population and at the individual patient level.

## VIII. Experimental Results

All models were evaluated using the same preprocessing pipeline and validation strategy to ensure a fair comparison. Performance was measured based on accuracy, F1-score, recall, and ROC–AUC, which are clinically important metrics for detecting gallstone cases.

The results showed meaningful variation between models, reflecting different strengths in sensitivity, generalization, and error distribution. While k-NN provided stable and interpretable results, tree-based models and the ensemble classifier achieved the best overall predictive performance.

The confusion matrices in Figure 8 illustrate how each model classified gallstone-positive and gallstone-negative patients. Random Forest, Gradient Boosting, and the ensemble model showed the most balanced performance, with low false negative rates. k-NN also performed well but produced slightly more false negatives. Logistic Regression and Naive Bayes struggled with detecting positive cases, leading to more misclassifications.

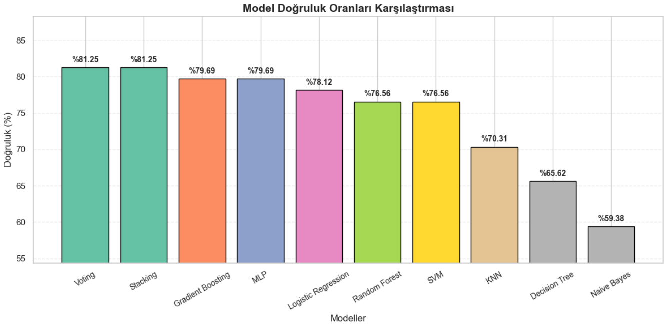

Fig. 7: Comparison of test accuracy across the evaluated machine learning models.

The results indicate that ensemble models effectively capture non-linear clinical patterns, while k-NN provides stable generalization. The high accuracy of tree-based methods suggests that interactions between metabolic markers and body composition are key for prediction. This supports the use of similarity-based models to identify ”look-alike” patient profiles in clinical decision support systems.

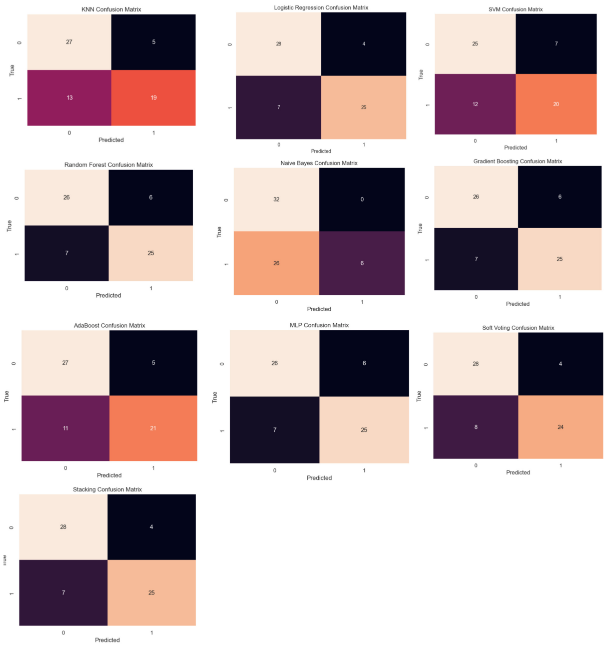

Fig. 8: Confusion Matrices of Evaluated Models.

Overall, the visualization demonstrates that most models successfully identified gallstone patients, but tree-based and ensemble methods provided the strongest diagnostic reliability.

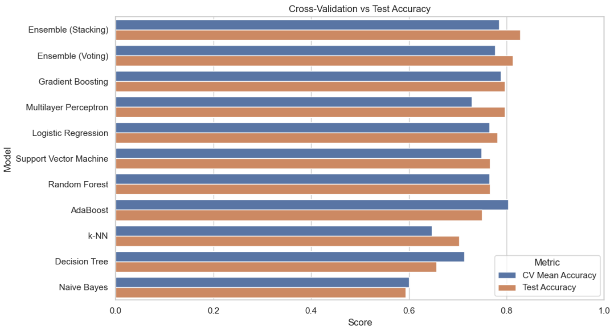

Fig. 9: Cross-Validation vs Test Accuracy

Figure 9 compares cross-validation and test accuracy across all models, showing strong consistency between the two. This indicates that the models generalize well, are not overfitted, and capture stable patterns within the data.

Ensemble, Gradient Boosting and Multilayer Perceptron models achieved the highest accuracy values, reflecting their robustness and suitability for complex clinical datasets. These results suggest strong potential for real-world decision support use.

Naive Bayes and Decision Tree models performed lower, consistent with their limitations in modeling nonlinear interactions.

The k-NN classifier showed closely aligned validation and test accuracy, demonstrating reliable generalization and strong compatibility with the dataset structure.

Overall, the results confirm that the evaluated models perform reliably on unseen data and support the validity of the proposed experimental framework.

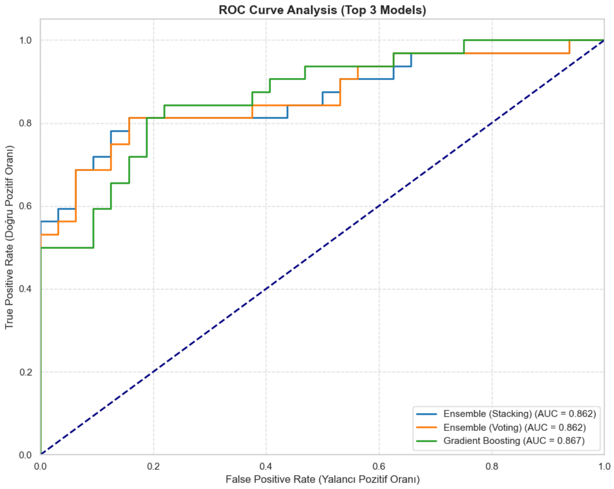

Fig. 10: ROC Curve Analysis (Top Models)

Figure 10 shows that Random Forest, Gradient Boosting, and Multilayer Perceptron effectively distinguish between positive and negative gallstone cases, outperforming the diagonal baseline. Gradient Boosting and Random Forest achieve higher true positive rates and smoother curves, while the Multilayer Perceptron performs competitively. Overall, the results confirm strong risk separation and support the models' potential for non-invasive decision support.

Figure 11 shows the global SHAP value distribution for the Random Forest model, highlighting the most influential features in gallstone prediction. The results indicate that multiple bioimpedance and laboratory variables contribute meaningfully to model output, supporting the multifactorial nature of gallstone risk.

Key predictors such as CRP, BM, ICW, LM, Gender, and ECF/TBW measures align with known metabolic and inflammatory pathways associated with gallstone formation. Higher adiposity-related values generally increased predicted risk, whereas reduced HDL and abnormal Vitamin D levels were linked to lower scores.

Additional contributors, including water distribution, bone mass, hemoglobin, and creatinine, suggest broader physiological involvement spanning hydration, body composition, and renal–hepatic function.

Overall, the SHAP results enhance interpretability by demonstrating that the model's feature importance patterns are

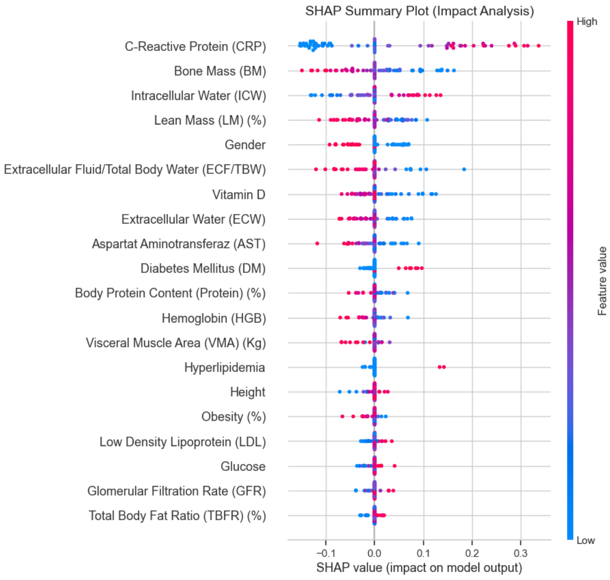

Fig. 11: SHAP Global Feature Impact (Beeswarm)

biologically plausible and clinically meaningful, reinforcing its suitability as a decision support tool.

## IX. Discussion

The findings of this study align with established metabolic, demographic, and biochemical risk patterns reported in the gallstone literature. Body composition indicators, particularly BMI and visceral fat metrics, were consistently influential across models, reinforcing the well-known association between adiposity, hepatic lipid imbalance, and biliary cholesterol supersaturation. Laboratory markers including CRP, glucose, lipid profile components, and Vitamin D also contributed meaningfully to predictions, reflecting inflammatory and metabolic pathways previously linked to gallstone pathophysiology.

Methodologically, ensemble models such as Random Forest and Gradient Boosting demonstrated the strongest overall performance, while Naive Bayes and Decision Tree models showed lower accuracy—consistent with the nonlinear and interdependent nature of clinical predictors. k-NN achieved stable generalization and strong interpretability, indicating that similarity-based learning aligns well with the clinical data structure and may offer practical value in patient-specific assessments.

SHAP analysis enhanced transparency by confirming that highly ranked features correspond to clinically plausible determinants of gallstone risk.

Overall, the results support the feasibility of using routine clinical measurements to predict gallstone risk without imaging data. Future work should expand cohort size, validate models externally, and incorporate longitudinal data to strengthen generalizability and clinical applicability.

## X. Conclusion and Future Work

This study demonstrates the feasibility of using routinely collected clinical, biochemical, and bioimpedance measurements to predict gallstone risk through machine learning. The results indicate that k-NN, supported by ensemble and boosted models, offers strong classification performance and interpretable similarity-based insights, highlighting its potential value in early risk stratification. Importantly, the proposed non-invasive approach reduces reliance on imaging-based diagnosis and supports data-driven clinical decision making.

Future work will prioritize external validation using larger and multi-center cohorts to enhance generalizability. Additional studies may explore model deployment within clinical workflows, real-time risk scoring, and integration with electronic health records. Further improvements through feature selection, longitudinal evaluation, and hybrid modeling approaches may also increase predictive accuracy and clinical applicability.

## Acknowledgments

This study was conducted within the scope of a TÜBİTAK 2209-A Undergraduate Research Project.

Github: https://github.com/huseyinky40/gallstone-prediction-knn-analysis

Dataset: https://archive.ics.uci.edu/dataset/1150/gallstone-1

## References

[1] W. Kratzer, R. Mason, and V. Kachele, “Prevalence of gallstones in sonographic surveys worldwide,” *Journal of Clinical Ultrasound*, vol. 27, pp. 1–7, 1999.

[2] A. Attili, N. Carulli, E. Roda, *et al.*, “Epidemiology of gallstone disease in italy,” *American Journal of Epidemiology*, vol. 141, pp. 158–165, 1995.

[3] J. Amaral and W. Thompson, “Gallbladder disease in the morbidly obese,” *American Journal of Surgery*, vol. 149, pp. 551–557, 1985.

[4] K. Heaton *et al.*, “Symptomatic and silent gall stones in the community,” *Gut*, vol. 32, pp. 316–320, 1991.

[5] U. Gustafsson *et al.*, “Changes in bile composition after bariatric surgery,” *Hepatology*, vol. 41, pp. 1322–1328, 2005.

[6] N. Mendez-Sanchez *et al.*, “Role of diet in cholesterol gallstone formation,” *Clinica Chimica Acta*, vol. 376, pp. 1–8, 2007.

[7] A. Colecchia *et al.*, “Predicting gallstone progression using motility,” *American Journal of Gastroenterology*, vol. 101, pp. 2576–2581, 2006.

[8] M. Stampfer *et al.*, “Risk of symptomatic gallstones in women with severe obesity,” *American Journal of Clinical Nutrition*, vol. 55, pp. 652–658, 1992.

[9] R. Eckel, S. Grundy, and P. Zimmet, “The metabolic syndrome,” *The Lancet*, vol. 365, pp. 1415–1428, 2005.

[10] F. Lammert and T. Sauerbruch, “Cholesterol gallstone disease overview,” *The Lancet*, vol. 368, pp. 230–239, 2006.

[11] F. Nervi *et al.*, “Influence of legume intake on biliary lipids,” *Gastroenterology*, vol. 96, pp. 825–830, 1989.

[12] C. Tsai *et al.*, “Glycemic load and gallstone disease,” *Gastroenterology*, vol. 129, pp. 105–112, 2005.

[13] R. Scragg *et al.*, “Plasma lipids and insulin in gall stone disease,” *British Medical Journal*, vol. 289, pp. 521–525, 1984.

[14] J. Everhart, M. Khare, M. Hill, and K. Maurer, “Prevalence and ethnic differences in gallbladder disease in the united states,” *Gastroenterology*, vol. 117, pp. 632–639, 1999.

[15] R. Sampliner, P. Bennett, L. Comess, F. Rose, and T. Burch, “Gallbladder disease in pima indians,” *New England Journal of Medicine*, vol. 283, pp. 1358–1364, 1970.

[16] F. Nervi *et al.*, “Gallbladder disease is associated with insulin resistance,” *Journal of Hepatology*, vol. 45, pp. 299–305, 2006.