Machine Learning for Gallstone Prediction

Adam Jeribi
Department of engineering Faculty of electrical engineering and computer science
University of Ottawa
Ottawa, Canada
ajeri020@uottawa.ca

Daniel Kurtz
Department of Neurosciences Faculty of Medicine
University of Ottawa
Ottawa, Canada
dkurt044@uottawa.ca

Othmane Daali
Department of engineering Faculty of electrical engineering and computer science
University of Ottawa
Ottawa, Canada
odaali@uottawa.ca

Vahid Kamandlooie
Department of engineering

Faculty of electrical engineering and computer science
University of Ottawa
Ottawa, Canada
vkama080@uottawa.ca

Group 19 – DTI 5126 / IAI 5120

***Abstract*— Gallstone disease is a highly prevalent hepatobiliary disorder affecting millions worldwide. This study presents a comprehensive machine learning approach for gallstone prediction using clinical, anthropometric, and laboratory measurements. We evaluated three supervised learning models: Logistic Regression with L2 Regularization, Random Forest Classifier, and Artificial Neural Networks alongside unsupervised K-Prototypes clustering analysis. After feature engineering and multicollinearity treatment, we achieved good predictive performance. K-Prototypes clustering identified two distinct patient subgroups with varying gallstone risk profiles, enabling targeted clinical interventions. This work demonstrates the efficacy of data-driven approaches in enhancing diagnostic accuracy and patient risk stratification for gallstone disease.**

Keywords*— *Gallstone disease, Machine learning, Deep learning, Feature engineering, Neural networks, Logistic regression, Random Forest Classifier, Clustering analysis

I. Executive Summary

Gallstone disease, also known as choledocholithiasis, is a common disease seen globally. Current diagnosis methods, including ultrasound, CT, and MRI, are costly and might not be accurate in some patients. In this project several machine learning algorithms were trained to predict gallstone status based on lab test results. These models could be used to help physicians assess the risk of developing the disease and intervene at an early stage.

The team used a dataset from the UCI Machine Learning Repository, including 320 patients, half gallstone positive and half negative. The dataset includes 38 features, including demographics (such as age and gender), body composition (such as BMI, visceral fat, and total body fat ratio), and laboratory values (such as glucose and cholesterol). Except for handling outliers, no missing data or inconsistencies were observed in the dataset. Some features, however, were correlated with each other. In feature engineering, this issue was solved by using domain knowledge to create new features and remove highly correlated features. Exploratory analysis showed that obesity-related measures, especially BMI, visceral fat area (VFA), and total body fat ratio, are among the most relevant features for distinguishing between patients with and without gallstones.

Supervised and unsupervised machine learning models were developed. For supervised learning, Logistic Regression and Random Forest classifiers were trained using robust scaling and simple pipelines. On a balanced test set, the best model achieved an accuracy of about 78% and an F1 score of 0.77, with reasonably balanced performance. The model was slightly less sensitive in detecting positive gallstone cases, but this could be improved by adjusting the decision threshold or using cost-sensitive training to prioritize reducing false negatives. Having more patient entries and features that are more predictive of the risk of developing gallstones, such as sex hormone levels, would have led to better results.

K-Means clustering was applied to the dataset as an unsupervised baseline to see whether patients can be naturally grouped into patterns related to gallstone status. The clustering showed only partial alignment with the true labels, confirming that unsupervised methods alone are not sufficient for reliable diagnosis, but they do reveal some structure, especially around gender and body composition.

Overall, the findings suggest that relatively simple supervised models, trained on standard clinical and bioimpedance data, can provide useful, interpretable predictions of gallstone status. Future work could explore more advanced methods such as gradient boosting and CNN, and model validation on larger and more diverse patient populations would improve generalizability and clinical impact.

II. Problem Framing

Gallstone disease, also known as choledocholithiasis, is a common disease seen globally. Due to changes in lifestyle in recent years, the number of gallstones (GS) has increased every year, becoming a major world health problem [1]. traditional diagnosis methods such as CT scan, MRI, and ultrasonography are costly and have some limitations in certain populations [2]. Machine learning allows scientists to integrate huge and diverse data sources and reveal unseen

patterns to diagnose this disease at an early stage and develop an optimal care plan [3]. Moreover, using machine learning for diagnosis can reduce the chance of incorrect medical diagnoses, which may arise due to clinician stress, fatigue, or inexperience [2].

As an exocrine organ which produces bile salts, the gallbladder has a critical role in managing fat absorption in the small intestine. In some patients, the fatty components of bile accumulate, forming crystals of various sizes [2]. These crystals, or "stones", can accumulate and block the passage of bile into the intestines, potentially resulting in pain or an infection [4]. This often necessitates the removal of the gallbladder, a major surgery which requires patients to make significant dietary changes [5]. As a result, it is imperative that a method be developed to predict whether patients have or do not have gallstones from their clinical data to prevent more severe symptoms and complications from developing.

After reviewing 1000 papers systematically in this domain, Ahmed et al concluded that machine learning models show promising performance in diagnosing gallstones, although reliability and generalizability still need to be focused and improved [6].

III. Dataset Description

The dataset used to create the models in this project is sourced from the University of California Irvine machine learning repository [7]. The dataset, consisting of clinical data, was collected from the Internal Medicine Outpatient Clinic of the Ankara VM Medical Park Hospital in Turkey. The dataset consists of 319 patients, of whom 161 were diagnosed with gallstone disease, with 38 features. The feature space includes demographic, bioimpedance, and laboratory data. The dataset is complete, with no missing values, and is balanced between patients with and without gallstones.

The feature space is composed of 7 categorical features, exhibiting unbalanced class distributions as seen in the following table:

TABLE I
Categorical Variables' Frequency Distribution

Variable	Category	Count
Gender	0	162
	1	157
Hyperlipidemia	0	311
	1	8
Comorbidity	0	217
	1	99
	2	1
	3	2
Coronary Artery Disease (CAD)	0	307
	1	12
Hypothyroidism	0	310
	1	9
Diabetes Mellitus (DM)	0	276
	1	43
Hepatic Fat Accumulation (HFA)	0	129
	1	41
	2	122
	3	26
	4	1

In the remaining 31 numerical features, several features display varying distributions, highlighting potential outliers that may carry clinical significance.

The following table reports the mean, standard deviation, and percentile values for all numerical features, providing an overview of their dispersion and typical ranges.

TABLE II
Numerical Variables' Descriptive statistics

Feature	**Mean**	**Std**	**50%**	**75%**	**Max**
Age	48.07	12.11	49.00	56.00	96.00
Comorbidity	0.335	0.517	1.00	1.000	3.00
Height	167.16	10.05	168.00	175.00	191.00
Weight	80.56	15.71	78.80	91.25	143.50
Body Mass Index (BMI)	28.88	5.31	28.30	31.85	49.70
Total Body Water (TBW)	40.59	7.93	39.80	47.00	66.20
Extracellular Water (ECW)	17.07	3.16	17.10	19.40	27.80
Intracellular Water (ICW)	23.61	5.35	23.00	27.55	52.00
Extracellular Fluid / TBW	42.21	3.24	42.00	44.00	57.00
Total Body Fat Ratio (%)	28.27	8.44	27.82	34.81	50.92
Lean Mass (%)	71.64	8.44	72.12	77.85	93.67
Body Protein (%)	15.94	2.33	15.87	17.43	24.81
Visceral Fat Rating (VFR)	9.08	4.33	9.00	12.00	31.00
Bone Mass (BM)	2.80	0.51	2.80	3.20	4.00
Muscle Mass (MM)	54.27	10.60	53.90	62.60	78.80
Obesity (%)	35.85	109.80	25.60	41.75	1954.00
Total Fat Content (TFC)	23.49	9.61	22.60	28.55	62.50

Visceral Fat Area (VFA)	12.17	5.26	11.59	15.10	41.10
Visceral Muscle Area (Kg)	30.40	4.46	30.41	33.80	41.10
Glucose	108.69	44.85	98.00	109.00	575.00
Total Cholesterol (TC)	203.50	45.76	198.00	233.00	360.00
Low Density Lipoprotein (LDL)	126.65	38.54	122.00	151.00	293.00
High Density Lipoprotein (HDL)	49.48	17.72	46.50	56.00	273.00
Triglyceride	144.50	97.90	119.00	172.00	838.00
AST	21.68	16.70	18.00	23.00	195.00
ALT	26.86	27.88	19.00	30.00	372.00
ALP	73.11	24.18	71.00	86.00	197.00
Creatinine	0.80	0.18	0.79	0.92	1.46
Glomerular Filtration Rate (GFR)	100.82	16.97	104.00	110.75	132.00
C-Reactive Protein (CRP)	1.85	4.99	0.22	1.62	43.40
Hemoglobin (HGB)	14.42	1.78	14.12	15.70	18.80
Vitamin D	21.40	9.98	22.00	28.06	53.10

The initial examination describing the overall feature distributions revealed substantial variability. This motivated conducting a more detailed analysis to further understand each attribute's behavior and clearly identify the outliers. To achieve this, boxplots were plotted for each numerical feature:

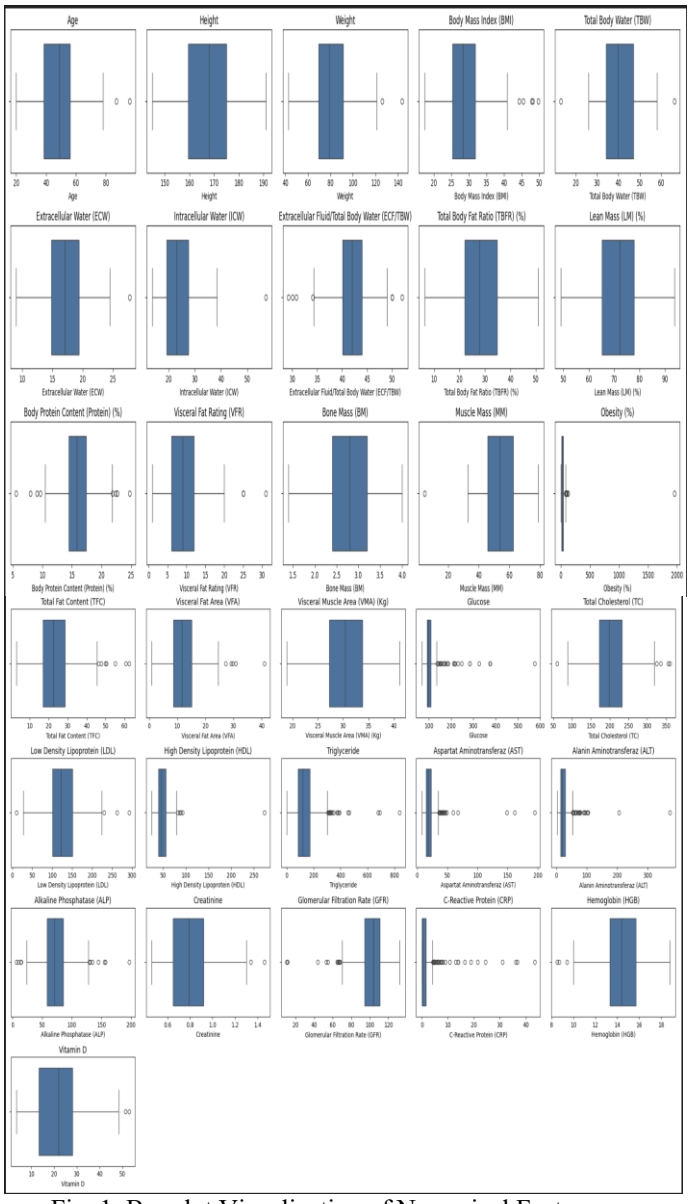

Fig. 1. Boxplot Visualization of Numerical Features

These boxplots helped identify two clear, medically impossible outliers, consistent with the values reported in Table I: 1954.00 in Obesity (%) and 273.00 in High Density Lipoprotein (HDL).

Nonetheless, additional outliers are present and are statistically notable; however, they were retained since they remain clinically coherent and may represent important indicators relevant to the study's objectives.

Given the large number of features, it is important to filter and prioritize the most relevant variables for further analysis. To achieve this, a heatmap of the correlation matrix was plotted. This helped identify the features that are most strongly correlated. These features can serve as potential predictors for the gallstone status target variable.

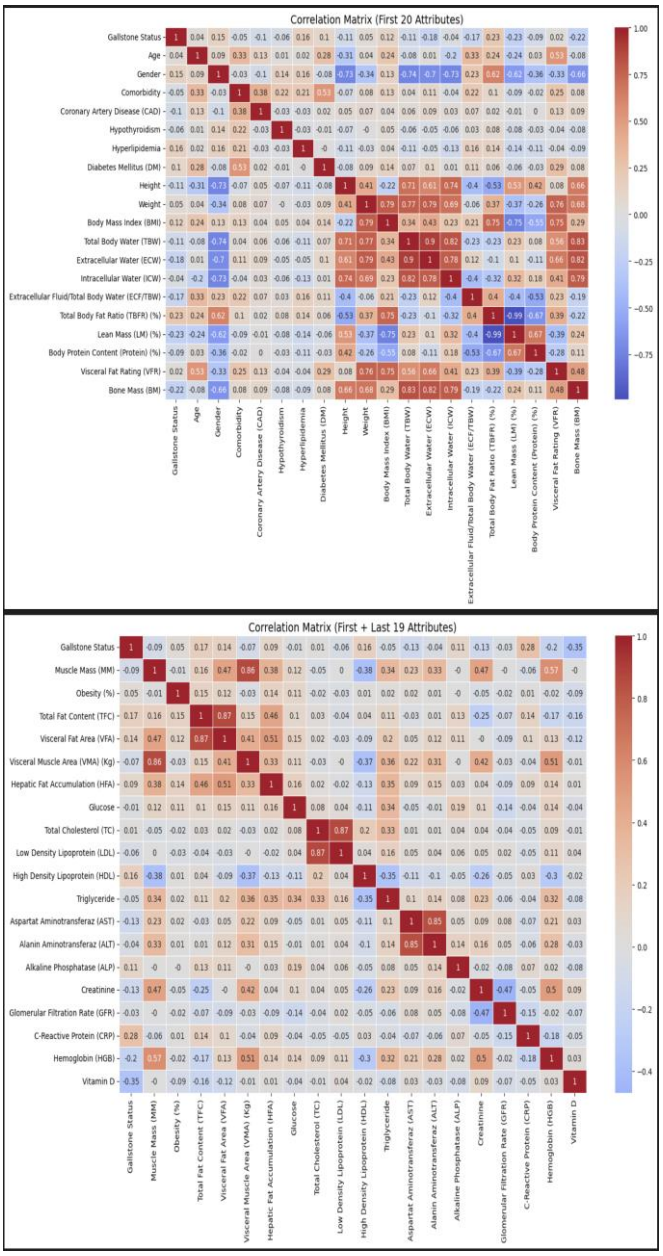

Fig. 2. Heatmap Correlation Matrix for all features

The 10 most correlated features were selected: Gender, Hyperlipidemia, Hepatic Fat Accumulation (HFA), Alkaline Phosphatase (ALP), Visceral Fat Area (VFA), High Density Lipoprotein (HDL), C-Reactive Protein (CRP), Total Body Fat Ratio (TBFR), Body Mass Index (BMI), Total Fat content (TFC).

These features showed a correlation greater than 0.1. But to avoid redundancy, two will be dropped: Body Mass Index (BMI) and Total Fat content (TFC). These two features describe nearly the same physiological phenomenon as VFA and TBFR, which will be kept for exhibiting a higher correlation with the target variable.

To further understand how these features impact the target variable, it is important to conduct a comparative analysis of how these features are distributed relatively to the gallstone status. A plot of histograms for numerical features was used to visualize the distribution of each feature. The following figure shows how elevated values of CRP are clear indicators of gallstones:

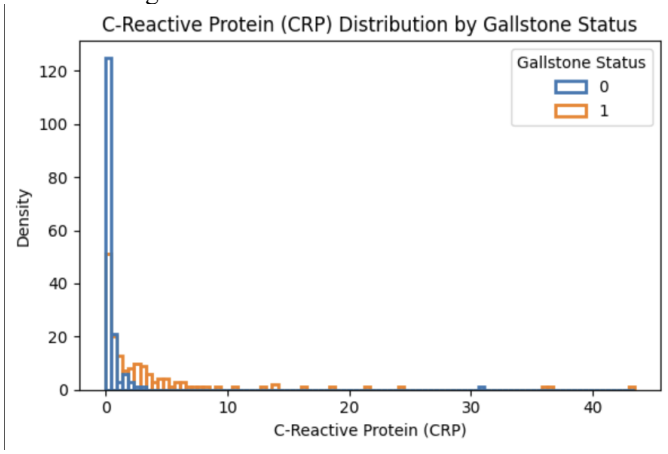

Fig. 3. Histogram of C-Reactive Protein (CRP) Distribution by Gallstone Status

For ALP, a higher value tends to indicate the presence of gallstones, but not as decisively as with CRP:

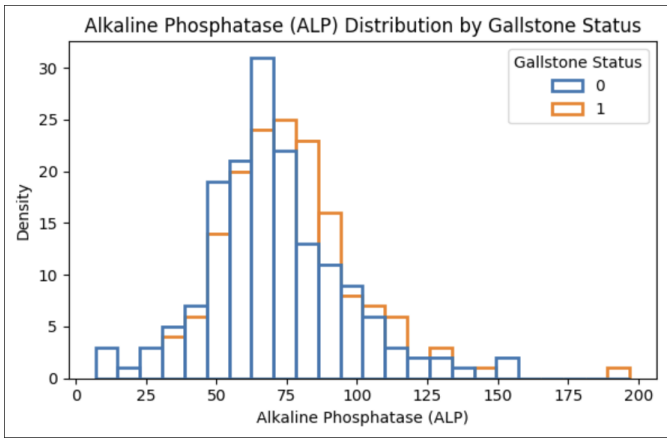

Fig. 4. Histogram of Alkaline Phosphatase (ALP) Distribution by Gallstone Status

For HDL, mid-to-high values can have more predictive importance:

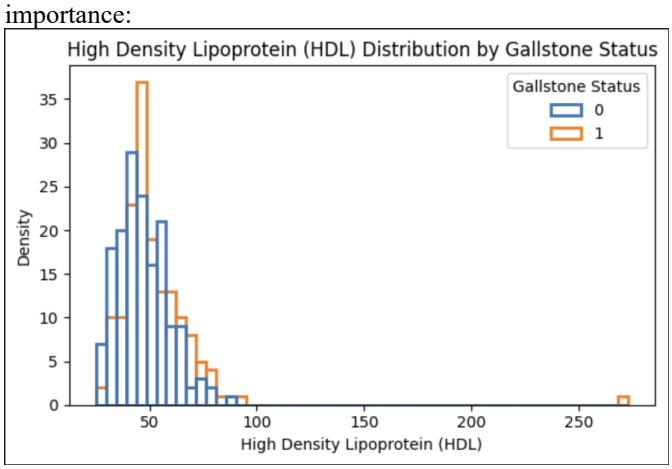

Fig. 5. Histogram of High-Density Lipoprotein (HDL) Distribution by Gallstone Status

A similar finding is seen for VFA:

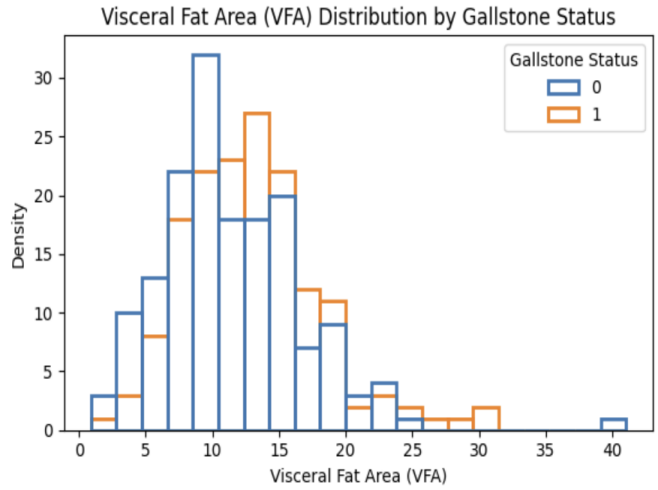

Fig. 6. Histogram of Visceral Fat Area (VFA) Distribution by Gallstone Status

Finally, a similar behavior can be noticed for TBFR:

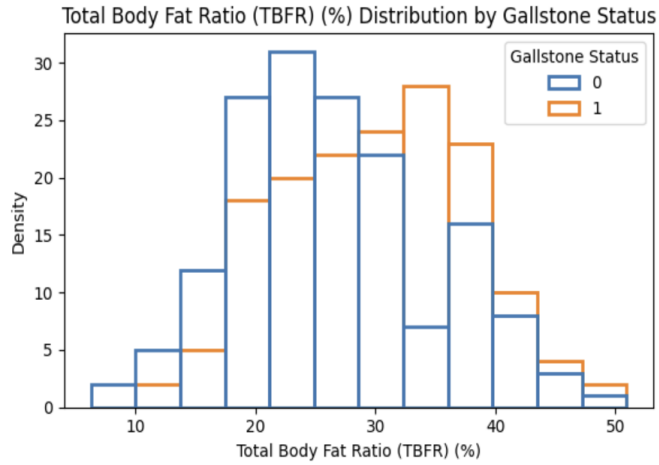

Fig. 7. Histogram of Total Body Fat Index (TBFR) Distribution by Gallstone Status

As for categorical attributes, bar charts are ideal for visualizing the relative distribution. The following bar chart shows how gallstone status varies across HFA categories, with moderate and high levels more associated with the presence of gallstones.

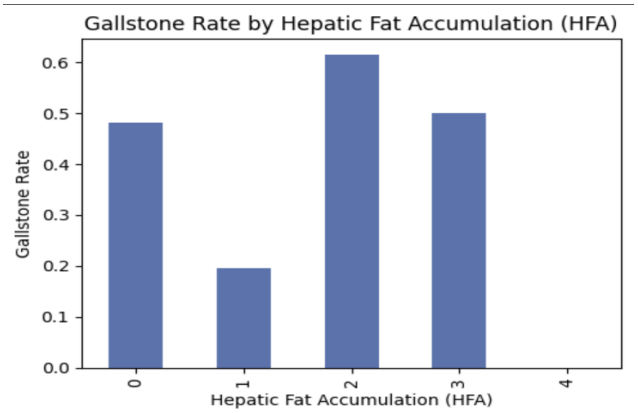

Fig. 8. Bar Chart of Hepatic Fact Accumulation (HFA) rate by Gallstone Status

Hyperlipidemia is another word for a high level of cholesterol in the blood, which is medically known to be a strong indicator for the presence of gallstones, as confirmed by this plot:

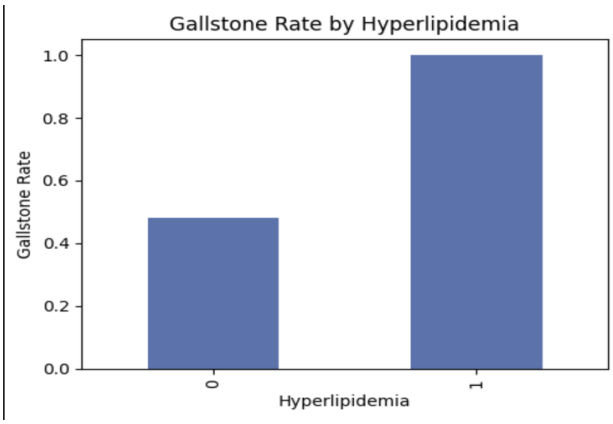

Fig. 9. Bar Chart of Hyperlipidemia rate by Gallstone Status

Reviewing the role of gender in gallstone disease, it is also known that female sex hormones, such as estrogen, can have more impact of the development of gallstones in a human body [4], as described in the following figure:

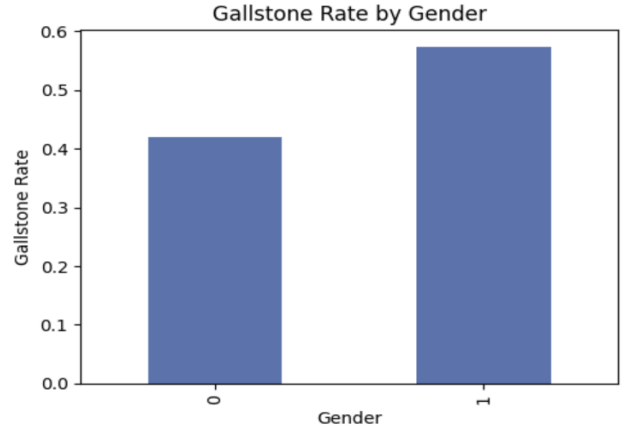

Fig. 10. Bar Chart of Gender rate by Gallstone Status

IV. Pre-processing

Pre-processing is an essential step in creating machine learning models due to the importance of preparing data for model training. Without correctly processed data, model inferences could be misleading or incorrect, reducing the potential accuracy of the models. To ensure that the data was correctly prepared prior to training, several methods were utilized.

Firstly, the dataset was checked for missing data or duplicate values to prevent pattern distortion or biases influencing training. As seen in Figure 11, no missing values were found.

```
# Checking for missing values in the dataset
print("Number of missing values:",df.isna().sum().sum())
```

```
[12]
Number of missing values: 0
```

Fig. 11. Code checking for missing values in the gallstone dataset, along with the corresponding result.

Additionally, as seen in Figure 12 below, no duplicate values were found, as the dataset as provided was complete with 319 unique patient entries.

```
# Checking for duplicates
num_duplicates = df.duplicated().sum()
df = df.drop_duplicates()
print("Number of duplicate rows :", num_duplicates)
```

```
Number of duplicate rows : 0
```

Fig. 12. Code checking for duplicate values in the gallstone dataset, along with the corresponding result.

If missing values existed within the dataset, they could have been handled by removing the affected rows or through data imputation. In data imputation, missing datapoints are filled in through a variety of approaches, such as using the mean, minimum, or maximum value of the column for numerical columns, or using the most inputted category for categorical columns. More advanced data imputation techniques, such as K-nearest neighbor (KNN) also exist for more complex datasets. If duplicate rows would have been found, the duplicates would have been removed.

Following detection and imputation of missing values and removal of duplicate values, the presence of outliers in the dataset was investigated. Outliers can skew dataset distributions and have an outsized impact on training behavior, distorting analysis. To identify outliers, the isolation forest technique, seen in Figure 13 below, was utilized, which identifies outlier rows.

```
dp = df.copy()
iso = IsolationForest(contamination=0.1, random_state=42)
dp['outlier_flag'] = iso.fit_predict(dp[numerical_attributes])
```

Fig. 13. Isolation forest code checking for outlier values in the gallstone dataset.

Following analysis of the data, approximately 10% of the dataset was flagged as outliers, as shown in Figure 14. However, as we are utilizing a medical dataset, with some patients' clinical measurements taken while they are in a medical crisis, some features would be expected to have a significant proportion of outliers. However, as seen in Data Visualization, there were two distinct outlier values in the Obesity Percentage and HDL columns which were medically impossible to attain. The two values were replaced with their columns' means for consistency.

```
outlier_count = (dp['outlier_flag'] == -1).sum()
normal_value = (dp['outlier_flag'] == 1).sum()

print(f"Outliers: {outlier_count}")
print(f"normal_value: {normal_value}")
```

```
[22] ✓ 0.0s
Outliers: 32
normal_value: 287
```

Fig. 14. Second half of isolation forest code checking for outlier values in the gallstone dataset, along with the corresponding results.

Following replacement of the two outlier values, scaling was completed to set all data scales from 0 to 1 and reduce the influence of the outliers. The RobustScaler technique, which scales data using the median and interquartile range, was used as it is less sensitive to outliers compared to the StandardScaler and MinMaxScaler techniques [8].

The categorical variables in the Gallstone Dataset were provided prior to processing in a numerical Boolean format, with values of zero or one, so categorical encoding was not necessary. However, if the dataset had categorical features which contained string datatype entries rather than numerical values, a OneHotEncoder could be used to convert the strings into encoded numerical values [9], a requirement for certain ML techniques, such as logistic regression.

Overall, compared to many similar datasets, this dataset was well-maintained, mostly accurate, and complete, so less pre-processing was required compared to other equivalent datasets.

V. Feature Extraction and Engineering

When preparing the dataset for training, feature multicollinearity was assessed. Based on the results of the multicollinearity analysis, 7 features with high correlation were dropped. These features were Extracellular Water (ECW), Intracellular Water (ICW), Total Body Fat Ratio (TBFR), Obesity, Total Fat Content (TFC), Lean Mass (LM), and Bone Mass (BM). Additionally, 21 new features were created to better reflect the team's needs based on domain knowledge and past usage in the prediction or prognostication of various pathologies. Firstly, BMI was changed from a continuous numerical variable into a binned categorical variable based on the WHO categories, while GFR was changed into a binned categorical variable based on the stages of Chronic Kidney Disease (CKD). An inflammation risk categorical variable was also created based on whether patients had a CRP value greater than 3. The other created variables are provided in the table below. These new features include body composition ratios, interaction between age and different variables, and patient comorbidity burden.

TABLE III
Created features and descriptions.

Feature	**Description (with Citation)**
TC_HDL_Ratio	Ratio of Total Cholesterol to HDL [10]
TG_HDL_Ratio	Ratio of Triglycerides to HDL [11]
LDL_HDL_Ratio	Ratio of LDL to HDL [12]
NonHDL_Cholesterol	Total Cholesterol – HDL [13]
AST_ALT_Ratio	Ratio of AST to ALT (De Ritis Ratio) [14]
Liver_Enzyme_Sum	Sum of AST and ALT [15]

High_TG	Boolean feature, either positive (TG > 150) or negative (TG <= 150)
Low_HDL	Boolean feature, either positive (HDL < 40 in males, HDL < 50 in females) or negative (HDL above those levels)
High_Glucose	Glucose > 100 mg/dl [16]
MetS_Score	Sum of High_TG, Low_HDL, and High_Glucose variable metrics [17]
Age_BMI_Interaction	Age * BMI
Age_Diabetes	Age * Diabetes [18]
Age_VFA	Age * VFA
VAI_Proxy	Proxy metric for VAI, equal to (VFA*Triglyceride)/HDL [19]
ViscMuscle_Ratio	Ratio of Visceral Fat Area (VFA) / Muscle Mass [20]
Fat_Muscle_Ratio	Ratio of Total Body Fat Ratio (TBFR) / Muscle Mass [21]
Comorbidity_Count	Sum of Boolean variables, Coronary Artery Disease (CAD) + Hypothyroidism + Hyperlipidemia + Diabetes [22]

VI. Models Tested

1) *Supervised models*

a) Logistic Regression:

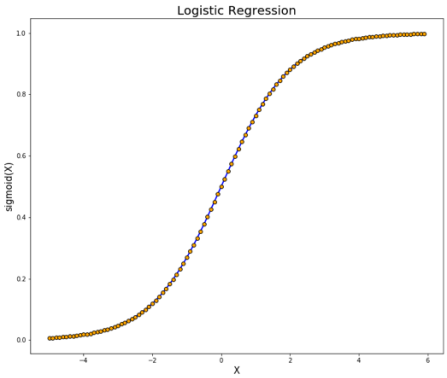

Fig. 15. Logistic Regression

Logistic Regression [23] is a supervised machine learning algorithm used for classification, often for binary classification tasks (0 or 1). An example of a logistic regression curve can be seen above in Figure 15.

When training the model, 5-fold stratified cross-validation was utilized on the 70% training subset to generate robust performance estimates. The model used the default parameters such as regularization strength at C=1.0 to provide a balance between regularization and model fit. The maximum number of iterations was set to 1000 to ensure convergence for this small-sized dataset.

One of the main advantages of logistic regression is its full interpretability. Each coefficient shows how a feature changes the log-odds of the outcome. In addition, it offers fast inference for real-time use, well calibrated probabilities for decision-making, and robustness to small model misspecifications.

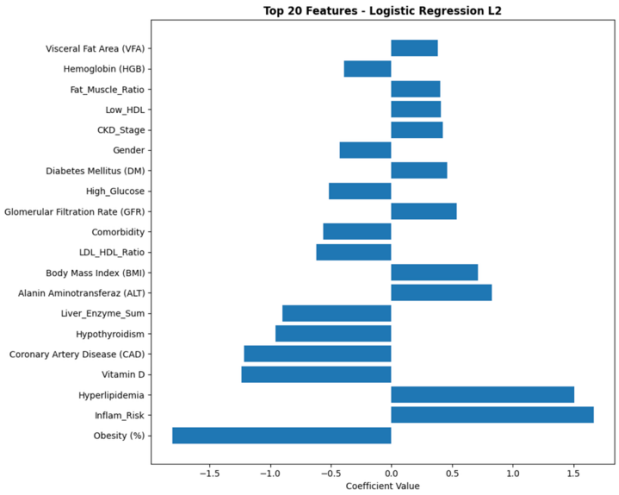

Fig. 16. Most important features for logistic regression

b) Random Forest Classifier:

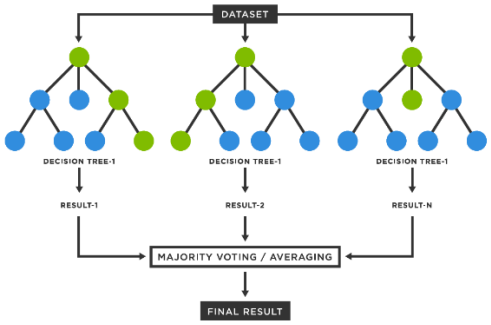

Fig. 17. Random Forest Classifier

The Random Forest Classifier technique creates multiple decision trees on different random subsets of the data and features, then combines their predictions, by using the majority vote approach, to produce a more accurate and stable classification [24]. This approach is visualized above in Figure 16.

The model used in the training utilized 200 individual decision trees; each trained on a random bootstrap sample of the training data with random feature subsets at each split. The depth of each tree was limited to 10 levels to prevent overfitting, the minimum number of samples required to split nodes was 10, and the minimum number of samples per leaf node was 5. When training the model, it used 5-fold stratified cross-validation, the same approach discussed for the logistic regression model.

The main advantage of Random Forest Classifier is the ability to capture non-linear feature relationships. Having multiple decision trees provides robustness through voting where individual tree errors average out. Furthermore, it is robust to outliers because individual splits are threshold-based rather than distance-based.

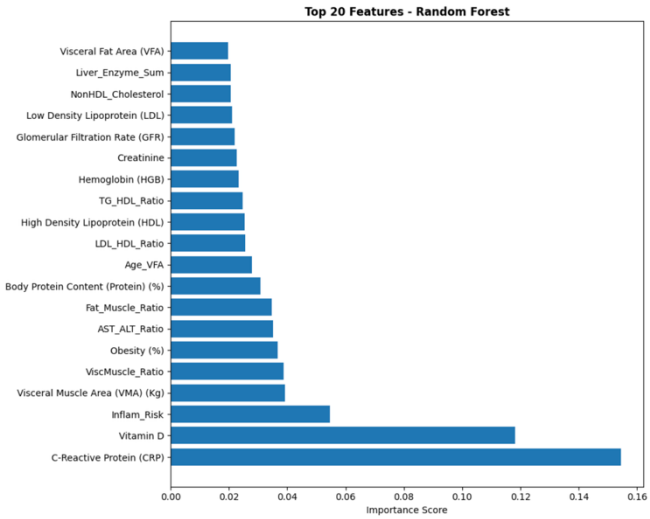

Fig. 18. Most important features for Random Forest

c) Neural Network:

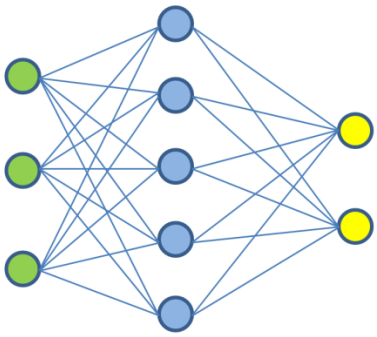

Fig. 19. Artificial Neural Network

The Artificial Neural Network (ANN) modelling technique was also used for this dataset. An ANN is a machine learning model inspired by the structure of the human brain. It consists of three different layers of interconnected nodes that mimic the brain neurons and learns patterns in the data by adjusting connection weights in the training with backpropagation, as shown in Figure 19 [25]. ANNs are mostly trained for deep learning purposes. That's why, they require huge amount of data to learn the overall dataset pattern and generalize well after being trained. Regardless, ANN was trained and tested for this dataset.

The architecture used for this shallow neural network was comprised of an input layer accepting 45 features, followed by a first hidden layer with 32 units and ReLU activation, a second hidden layer with 16 units and ReLU activation, and an output layer with 2 units (one per class) and a Softmax activation function. In the ANN, L2 regularization with $\lambda = 0.001$ was applied to all dense layers in order to prevent overfitting. It has also used Adam optimizer and progressed through only 15 epochs during training to prevent overfitting. A validation split of 15% was used to monitor performance on unseen data and to trigger early stopping.

Neural Networks are known for their capacity to learn complex non-linear data structures and patterns. However, due to the lack of data inside our dataset with only 319 records, the model wasn't capable of precisely detecting the overall pattern. Thus, this approach was only used for comparison with Logistic regression and Random Forest Classifier models.

2) Unsupervised models

a) K-prototypes Clustering:

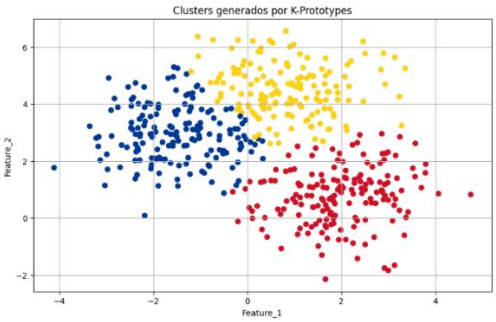

Fig. 20. K-prototypes clustering

K-prototypes clustering is a method designed for datasets that have both numerical and categorical features. It works by combining both K-means and K-modes [26]. For instance, numerical values are compared using regular distances, while categorical values are matched based on whether they are the same or different. A weighting factor keeps the two types of data balanced. An example of a PCA plot for a three-cluster K-prototypes model is shown above in Figure 20.

The primary advantage of K-Prototypes is the ability to handle mixed data types without converting categorical features to numerical features that would impact the distance calculations making it a good choice for this dataset.

Prior to training the K-prototypes model, the categorical and numerical features were identified, with 39 numerical features and 13 categorical features. Following setup, an elbow plot and silhouette analysis were completed, with the elbow plot shown below in Figure 21. The elbow plot does not possess a discernable slope decrease or levelling at any point from K=2 to K=6, so the plot does not settle on a specific optimal K value.

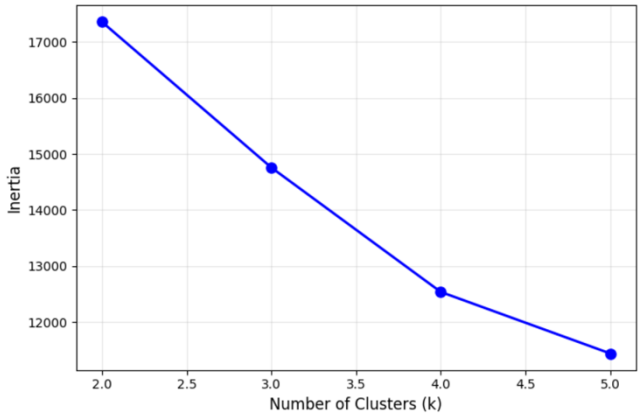

Fig. 21. Elbow plot for K-prototypes clustering model.

Unlike the elbow plot, the silhouette plot, shown in Figure 22, displays a maximum silhouette value of 0.84 at K=2, providing an optimal K-value to use in creating the models. This optimal K-value makes logical sense, as the unsupervised model was provided with the entire dataset and was able to identify the clear divide between patients with and without gallstones.

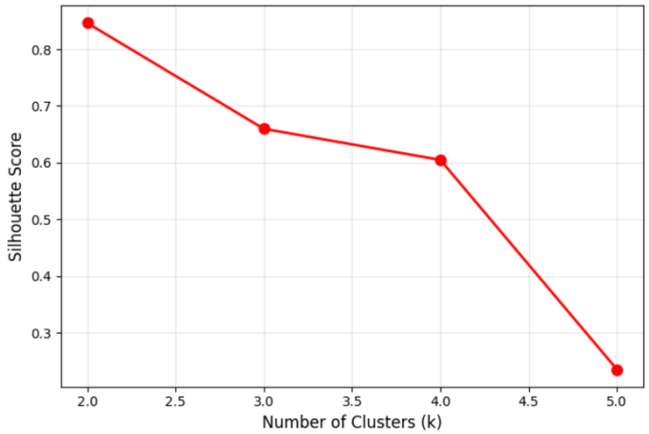

Fig. 22. Silhouette plot for K-prototypes clustering model.

Following elbow and silhouette analysis, the model was trained to cluster the dataset into two clusters. The distribution of patients with and without gallstones is provided below. As seen in the plot, cluster 0 possesses more patients with gallstones, while cluster 1 has a roughly equal distribution of patients with and without gallstones.

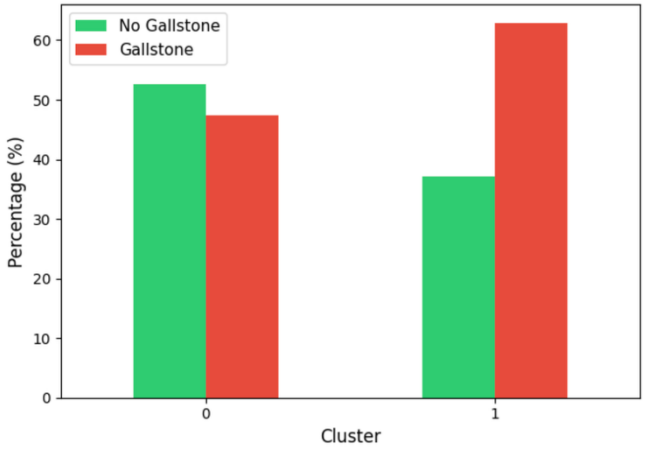

Fig. 23. Distribution of patients with and without gallstones in the two clusters produced by the unsupervised K-prototypes model.

When the two clusters were analyzed, there were clear differences, as seen below in Table IV. Firstly, all patients in cluster 0 were clinically diagnosed with diabetes, while none of the patients in cluster 1 had diabetes, indicating that this was the main differentiating feature the model used to group patients. This can also be seen via the Mean Blood Glucose measurement where the average in cluster 0 is 158.9 mg/dl compared to 100.9 mg/dl in cluster 1. Values of 70 to 99 mg/dl indicate a patient is healthy, values between 100 to 125 mg/dl indicate a patient has prediabetes, and values over 125 mg/dl are seen in patients with Diabetes Mellitus [27]. Therefore, while none of the patients in cluster 1 have diabetes, some are at risk of developing the disease in the future. Additionally, patients in cluster 0 were on average older, more obese, and had higher levels of triglycerides in their blood. Interestingly, despite having a higher average BMI and triglyceride score, the cluster 0 did not have a higher prevalence of hyperlipidemia compared to cluster 1.

TABLE IV
Mean values of patients in K-prototypes-produced cluster 0 and cluster 1.

	Cluster 0	**Cluster 1**
Mean Age	56.6	46.7
Mean BMI	30.8	28.6
Mean Visceral Fat Area (VFA)	13.8	11.9
Mean Blood Triglycerides (mg/dl)	196.1	136.5
Mean Blood Glucose (mg/dl)	158.9	100.9
Diabetes Prevalence	100%	0%
Hyperlipidemia Prevalence	2.30%	2.50%

The PCA plot for the clustering model is seen below in Figure 24. There are two clear groupings of patients seen, along a vertical line at the left of the plot, and towards the right side of the plot. Despite this distribution, patients with and without gallstones are seen in both clusters.

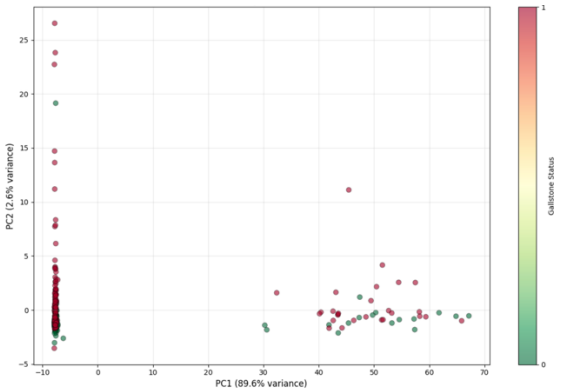

Fig. 24. Principal Component Analysis (PCA) plot for K-prototypes clustering model.

VII. Evaluation Results

1) *Logistic Regression*

Classification Report – Logistic Regression (TEST):

	precision	recall	f1-score	support
No Gallstone	0.76	0.73	0.74	22
Gallstone	0.78	0.81	0.79	26
accuracy			0.77	48
macro avg	0.77	0.77	0.77	48
weighted avg	0.77	0.77	0.77	48

Fig. 25. Logistic regression results

Logistic Regression performed well on the test set, reaching an overall accuracy of 0.77. The model showed a balanced performance between the two classes, with f1-scores of 0.74 for a "No Gallstone" status and 0.79 for a "Gallstone." status. Its recall was slightly higher for detecting gallstones sitting at 0.81, which means it is better at correctly identifying true positives. Overall, the model was consistent across precision, recall, and f1-score, proving it as a reliable classifier for this dataset.

Fig. 26. Logistic Regression confusion matrix

Using the testing dataset, which is a 15% subset from the original with only 48 records, the logistic regression model was able to correctly predict 16 true negatives and 21 true positives totaling to a solid 37 correct predictions out of 48.

2) *Random Forest Classifier*

Classification Report – Random Forest (TEST):

	precision	recall	f1-score	support
No Gallstone	0.71	0.77	0.74	22
Gallstone	0.79	0.73	0.76	26
accuracy			0.75	48
macro avg	0.75	0.75	0.75	48
weighted avg	0.75	0.75	0.75	48

Fig. 27. Random Forest Classifier results

The Random Forest Classifier model achieved solid performance with an accuracy score of 0.75, close to logistic regression. It performed in a similar manner across classes, with f1-scores of 0.74 for "No Gallstone" and 0.76 for "Gallstone." Unlike the neural network, the Random Forest Classifier captures useful patterns and maintains balanced precision and recall. Its performance was slightly lower than logistic regression, making it a robust model for classifying the data.

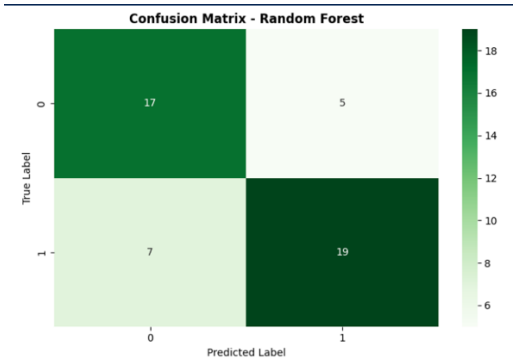

Fig. 28. Random Forest Classifier confusion matrix

For Random Forest Classifier, the results were almost the same as logistic regression. The model was able to correctly predict 17 true negatives and 19 true positives totaling to a reasonably accurate 36 correct predictions out of 48.

3) *Artificial Neural Network*

Classification Report – Neural Network (TEST):

	precision	recall	f1-score	support
...				
accuracy			0.33	48
macro avg	0.31	0.32	0.31	48
weighted avg	0.32	0.33	0.32	48

Fig. 29. Artificial Neural Network results

The Artificial Neural Network performed significantly worse than the two previous models, achieving only 0.33 accuracy. Precision, recall, and F1-scores are all around that range as well, indicating the model struggled to learn meaningful patterns from the data. This poor performance was expected due to the lack of records in the dataset (Only 319 records / 223 if we count only the training set), as ANNs are best suited to large datasets.

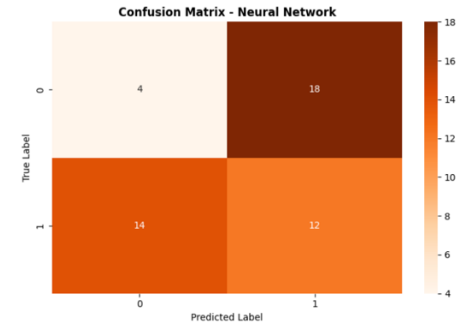

Fig. 30. ANN confusion matrix

As previously discussed, ANN had a poor performance which explains having many of both false negatives (14) and false positives (18). In addition, it was able to predict only 4 true negatives compared to the 12 true positives which suggests that it struggles especially with detecting a situation where a patient doesn't have "Gallstone" which is better than if it was the other way around.

VIII. Model Comparision

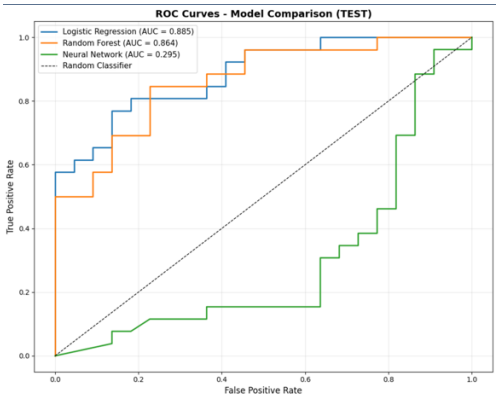

Fig. 31. ROC curve

The ROC curves, seen above indicate the ability of each model to separate the two classes at different thresholds.

As shown in figure 31, Logistic Regression performed the best overall, with the highest AUC value out of all models, 0.885, which means that it consistently keeps a high true-positive rate while keeping the false-positive rate low.

The Random Forest Classifier performed similarly, with an AUC of 0.864, showing strong discrimination but slightly less smooth performance at low thresholds. The ANN, however, performed very poorly. Its AUC was only 0.295. This clearly indicates that the neural network has failed to learn meaningful patterns in the data.

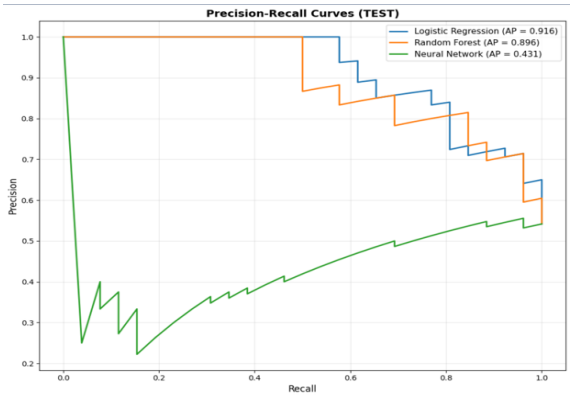

Fig. 32. Precision-Recall curve

A Precision-Recall curve, as seen above in Figure 32, was used to evaluate the performance of binary classification models such as the supervised models in this deliverable. Logistic Regression had the best performance, with an AP of 0.916, indicating that it maintains a high precision rate while still detecting most positive cases. The Random Forest Classifier performed almost as well, with an AP of 0.896. Finaly, ANN on the other hand, had a curve that was significantly lower, with an AP of just 0.431. This indicates that it produces many false positives and struggles to consistently predict the correct class.

Model Performance Comparison:

Model	CV_Accuracy (train CV)	CV_ROC-AUC (train CV)	Test_Accuracy	Test_Recall	Test_F1	Test_ROC-AUC	Test_PR-AUC
Logistic Regression L2	0.712828	0.764103	0.770833	0.807692	0.792453	0.884615	0.915765
Random Forest	0.708485	0.780345	0.750000	0.730769	0.760000	0.863636	0.895674
Neural Network	0.493232	NaN	0.333333	0.461538	0.428571	0.295455	0.431428

Fig. 33. Model Performance comparison

As shown in Figure 33, the Logistic Regression model using L2 regularization clearly outperformed both Random Forest Classifier and the ANN. It achieved the highest test accuracy with 0.771, a f1-score of 0.792, a ROC-AUC of 0.885, and a PR-AUC of 0.916, with a strong recall value of 0.808. It also showed very good performance at generalizing to the testing dataset. This is shown by having an accuracy score of 0.713 on the training subset and a 0.771 score on the testing subset. The Random Forest model performed well, but noticeably worse than logistic regression, especially in recall (0.731 for RF vs 0.808 for LR). This difference is quite significant, as detecting true positives, patients with the pathology, is a priority for this dataset. The Neural Network performed poorly across nearly all metrics (test accuracy of 0.33, recall = 0.462 and f1-score = 0.429).

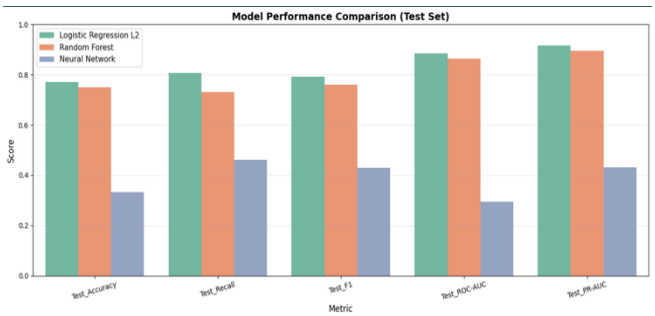

Fig. 34. Model performance visualization

To summarize, for this dataset, logistic regression was the most effective and reliable model, demonstrating that complex models like neural networks can fail to learn meaningful patterns in the dataset when the quantity of data is limited or imbalanced, while simpler models with proper parameter tuning generalize and perform better.

IX. Discussion

1) Key findings:

a) Predictive model performance:

The supervised models showed different levels of predictive ability, which reflects the varied and non-linear characteristics of gallstone disease. Logistic Regression with L2 regularization had the best performance, while the Random Forest Classifier performed almost as well but had slightly lower recall. The neural network did not perform well, mainly due to dataset limitations. This confirms that deep learning models need much larger datasets to work effectively. These findings suggest that in a clinical setting, simpler regularized models offer more reliable differentiation than complex structures.

b) Feature importance insights:

Across all models, consistent predictors such as visceral fat area (VFA), markers of metabolic issues, and C-reactive protein (CRP) are involved in gallstone disease. VFA stands out as the strongest predictor, showing that how fat is distributed is more important than overall body fat, confirming the team's hypotheses from the data description step of the project. Triglycerides, HDL, and glucose also score high in importance, and their combined metabolic syndrome score has even greater predictive power, indicating a combined metabolic risk. CRP underlines the role of systemic inflammation as a separate factor in gallstone development. Lastly, age and VFA interaction terms reveal that visceral fat becomes more harmful as people get older, showing how metabolic risk increases with age.

c) Multicollinearity management success

The multicollinearity management strategy effectively reduced feature redundancy by removing 18% of the variables while keeping, and in some cases improving, model performance. This indicates that the excluded features added noise instead of useful predictive information. It supports the use of correlation analysis as done in data description, VIF filtering, and clinical judgment together. Notably, clinically relevant variables like visceral fat area, BMI, liver enzymes, and lipid components were kept despite their inter-correlations. Removing them would have resulted in the loss of important mechanistic information. Regularization in logistic regression, ensemble averaging in Random Forest, and neural network constraints worked together to stabilize the models, showing that multicollinearity was effectively managed in feature selection and model-level regularization.

d) Clustering insights: Patient Phenotyping

K-Prototypes clustering was used on the mixed-type dataset after identifying 39 numerical and 13 categorical variables. Silhouette analysis showed that the best solution was K = 2, which matched the natural separation in metabolic profiles. The resulting clusters showed distinct phenotypes. Cluster 0 included older, more obese patients with much higher triglycerides, increased glucose levels, and a 100% diabetes rate. In contrast, Cluster 1 included younger individuals with healthier metabolic profiles and no diagnosis of diabetes. Even with these differences, gallstone-positive cases appeared in both clusters, showing that metabolic status strongly influences cluster membership, but does not solely determine it. The PCA projection also confirmed two clear patient groups, showing that the K-Prototypes model identified important patterns in the dataset.

e) Next steps

Additional machine learning models, such as xgBoost or other gradient-boosting tree techniques, could be investigated to see if the predictive accuracy increases further. While the dataset is small, a more complex neural network approach could be implemented to compare against the K-means model to determine which model is able to glean more insights from the dataset.

Neural networks, especially CNN, have proved to be outstanding in this domain [6], and in further steps, training proven algorithms along with new algorithms which have not been employed yet could be investigated.

X. Conclusion

This study created a complete machine learning pipeline for predicting gallstones. It included exploratory analysis, data cleaning, feature engineering, supervised modeling, and unsupervised clustering. The results show that understanding the clinical context is vital in data science. Domain knowledge helped with interpreting outliers, selecting features, and managing multicollinearity. Logistic Regression with L2 regularization delivered the best and most consistent predictive results. Random Forest was effective at identifying useful non-linear patterns. However, neural networks did not perform well due to the small size of clinical datasets. This highlights the need for more samples in deep learning applications.

Comparing our results with the paper that was published on this dataset [7], shows that the best model in this deliverable produced results which would be scientifically publishable. The table below shows a comparison between our results and the results of Esen et al.

TABLE V
Comparison of results between this deliverable and Esen et al.

Model	Precision		Recall		F1		AUC	
	Group 19	Esen et al	Group 19	Esen et al	Group 19	Esen et al	Group 19	Esen et al
Logistic regression	0.78	0.86	0.81	0.83	0.79	0.83	0.92	0.84
Random forest	0.79	0.91	0.73	0.81	0.76	0.86	0.89	0.85
Neural network	0.4		0.46		0.42		0.43	

GB		0.91		0.81		0.86		0.85

Unsupervised K-Prototypes clustering identified two patient groups that matter clinically, mainly based on metabolic status. This shows how unsupervised learning can assist in risk assessment without labeled outcomes.

Overall, the findings stress that thorough preprocessing, careful model choice, and interpretation based on domain knowledge are essential for getting reliable insights from clinical data and developing better predictive tools for gallstone disease.

REFERENCES

[1] L. Deng *et al.*, "Relative Fat Mass and Physical Indices as Predictors of Gallstone Formation: Insights From Machine Learning and Logistic Regression," *Int. J. Gen. Med.*, vol. 18, pp. 509–527, Jan. 2025, doi: 10.2147/IJGM.S507013.

[2] İ. Esen, H. Arslan, S. Aktürk Esen, M. Gülşen, N. Kültekin, and O. Özdemir, "Early prediction of gallstone disease with a machine learning-based method from bioimpedance and laboratory data," *Medicine (Baltimore)*, vol. 103, no. 8, p. e37258, Feb. 2024, doi: 10.1097/MD.0000000000037258.

[3] N. M. Salem *et al.*, "Machine and deep learning identified metabolites and clinical features associated with gallstone disease," *Comput. Methods Programs Biomed. Update*, vol. 3, p. 100106, Jan. 2023, doi: 10.1016/j.cmpbup.2023.100106.

[4] M. W. Jones, C. B. Weir, and M. Marietta, "Gallstones (Cholelithiasis)," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2025. Accessed: Nov. 30, 2025. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK459370/

[5] "Gallstones | Liver Canada." Accessed: Nov. 30, 2025. [Online]. Available: https://liver.ca/gallstones/

[6] A. S. Ahmed *et al.*, "Advancements in Cholelithiasis Diagnosis: A Systematic Review of Machine Learning Applications in Imaging Analysis," *Cureus*, vol. 16, no. 8, Aug. 2024, doi: 10.7759/cureus.66453.

[7] İ. Esen, H. Arslan, S. Aktürk Esen, M. Gülşen, N. Kültekin, and O. Özdemir, "Early prediction of gallstone disease with a machine learning-based method from bioimpedance and laboratory data," *Medicine (Baltimore)*, vol. 103, no. 8, p. e37258, Feb. 2024, doi: 10.1097/MD.0000000000037258.

[8] "StandardScaler, MinMaxScaler and RobustScaler techniques - ML," GeeksforGeeks. Accessed: Nov. 29, 2025. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/standardscaler-minmaxscaler-and-robustscaler-techniques-ml/

[9] "OneHotEncoder," scikit-learn. Accessed: Nov. 29, 2025. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

[10] D. Zhou, X. Liu, K. Lo, Y. Huang, and Y. Feng, "The effect of total cholesterol/high-density lipoprotein cholesterol ratio on mortality risk in the general population," *Front. Endocrinol.*, vol. 13, p. 1012383, Dec. 2022, doi: 10.3389/fendo.2022.1012383.

[11] P. Baneu *et al.*, "The Triglyceride/HDL Ratio as a Surrogate Biomarker for Insulin Resistance," *Biomedicines*, vol. 12, no. 7, July 2024, doi: 10.3390/biomedicines12071493.

[12] E. Davis, F. Huffman, and E. Onuoha, "Is the LDL/HDL- Cholesterol Ratio a Better Risk Indicator for Coronary Heart Disease Than the TC/HDL-Cholesterol Ratio?," *FASEB J.*, vol. 29, no. S1, p. 898.48, 2015, doi: 10.1096/fasebj.29.1_supplement.898.48.

[13] V. Raja *et al.*, "Non-HDL-cholesterol in dyslipidemia: Review of the state-of-the-art literature and outlook," *Atherosclerosis*, vol. 383, p. 117312, Oct. 2023, doi: 10.1016/j.atherosclerosis.2023.117312.

[14] M. Botros and K. A. Sikaris, "The De Ritis Ratio: The Test of Time," *Clin. Biochem. Rev.*, vol. 34, no. 3, pp. 117–130, Nov. 2013.

[15] H. O. Abdallah, M. F. Weingart, R. Fuller, D. Pegues, R. Fitzpatrick, and B. J. Kelly, "Subglottic suction frequency and adverse ventilator-associated events during critical illness," *Infect. Control Hosp. Epidemiol.*, vol. 42, no. 7, pp. 826–832, July 2021, doi: 10.1017/ice.2020.1298.

[16] "What Does My Blood Glucose Test Result Mean?," Cleveland Clinic. Accessed: Nov. 30, 2025. [Online]. Available: https://my.clevelandclinic.org/health/diagnostics/12363-blood-glucose-test

[17] H. Chen, S. Peng, R. A, M. Chen, L. Yuan, and M. Long, "Diagnostic value and mediation effects of the visceral adiposity index, triglyceride-glucose index, and platelet-to-HDL ratio in young overweight and obese Chinese adults," *Front. Nutr.*, vol. 12, Sept. 2025, doi: 10.3389/fnut.2025.1599603.

[18] C. T. Cigolle, C. S. Blaum, C. Lyu, J. Ha, M. Kabeto, and J. Zhong, "Associations of Age at Diagnosis and Duration of Diabetes With Morbidity and Mortality Among Older Adults," *JAMA Netw. Open*, vol. 5, no. 9, p. e2232766, Sept. 2022, doi: 10.1001/jamanetworkopen.2022.32766.

[19] M. C. Amato *et al.*, "Visceral Adiposity Index," *Diabetes Care*, vol. 33, no. 4, pp. 920–922, Apr. 2010, doi: 10.2337/dc09-1825.

[20] Y. Cho *et al.*, "Skeletal muscle mass to visceral fat area ratio as a predictor of NAFLD in lean and overweight men and women with effect modification by sex," *Hepatol. Commun.*, vol. 6, no. 9, pp. 2238–2252, Sept. 2022, doi: 10.1002/hep4.1975.

[21] D. Liu, J. Zhong, Y. Ruan, Z. Zhang, J. Sun, and H. Chen, "The association between fat-to-muscle ratio and metabolic disorders in type 2 diabetes," *Diabetol. Metab. Syndr.*, vol. 13, no. 1, p. 129, Nov. 2021, doi: 10.1186/s13098-021-00748-y.

[22] J. Wang *et al.*, "Advancing neuroprotection and atherosclerosis prevention through familial hypercholesterolemia management: Analyzing comorbidity burden with stroke, coronary heart disease, hypertension, and diabetes," *Neuroprotection*, vol. 03, no. 01, pp. 95–103, Mar. 2025, doi: 10.1002/nep3.72.

[23] "What Is Logistic Regression? | IBM." Accessed: Nov. 29, 2025. [Online]. Available: https://www.ibm.com/think/topics/logistic-regression

[24] "What Is Random Forest? | IBM." Accessed: Nov. 29, 2025. [Online]. Available: https://www.ibm.com/think/topics/random-forest

[25] "What Is a Neural Network? | IBM." Accessed: Nov. 29, 2025. [Online]. Available: https://www.ibm.com/think/topics/neural-networks