ANKARA YILDIRIM BEYAZIT UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES



GALLSTONE DETECTION USING MACHINE LEARNING ALGORITHMS WITH METAHEURISTICS

M.Sc. Thesis by

Cansu Çalışkan Tamer

Department of Computer Engineering

April, 2025

ANKARA

GALLSTONE DETECTION USING MACHINE LEARNING ALGORITHMS WITH METAHEURISTICS

A Thesis Submitted to the

Graduate School of Natural And Applied Sciences of

Ankara Yildirim Beyazit University

In Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Engineering, Department of Computer Engineering

by

Cansu Çalışkan Tamer

April, 2025

ANKARA

M.Sc. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled "**GALLSTONE DETECTION USING MACHINE LEARNING ALGORITHMS WITH METAHEURISTICS**" completed by **CANSU ÇALIŞKAN TAMER** under supervision of **ASSOC. PROF. DR. HİLAL ARSLAN** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. Hilal ARSLAN

Supervisor

Asst. Prof. Dr. Tuba ÇAĞLIKANTAR

Jury Member

Asst. Prof. Dr. Fatma KÜÇÜK

Jury Member

Prof. Dr. İlyas ÇANKAYA

Director

Graduate School of Natural and Applied Sciences

ETHICAL DECLARATION

I hereby declare that, in this thesis which has been prepared in accordance with the Thesis Writing Manual of Graduate School of Natural and Applied Sciences,

- All data, information and documents are obtained in the framework of academic and ethical rules,
- All information, documents and assessments are presented in accordance with scientific ethics and morals,
- All the materials that have been utilized are fully cited and referenced,
- No change has been made on the utilized materials,
- All the works presented are original,

and in any contrary case of above statements, I accept to renounce all my legal rights.

Date: 2025, 18 April **Signature:** _____

Name & Surname: Cansu Çalışkan Tamer

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Assoc. Prof. HİLAL ARSLAN for invaluable support of my master's study and research, for her constant support, guidance, and encouragement throughout all stages of this thesis. It was a great honor to work with her. I am very grateful that she has been with me with great patience and understanding at every stage of my difficulties. At every stage of our study, she always encouraged and inspired me with the question of how we could do better. In this regard, I believe that I have conducted this study successfully under her guidance.

In addition to my advisor, I am also grateful for my thesis committee members, Asst. Prof. Dr. Tuba Çağlıkantar, and Asst. Prof. Dr. Fatma Küçük for their precious discussions and suggestions about my thesis.

Finally, I would like to extend my sincere thanks to my mother, Serap Çalışkan, who has supported me unconditionally throughout my life. I would also like to thank my husband Yılmaz Tamer, who has always been by my side and supported me throughout this process.

April, 2025 **Cansu Çalışkan Tamer**

GALLSTONE DETECTION USING MACHINE LEARNING ALGORITHMS WITH METAHEURISTICS

ABSTRACT

Gallstone disease is a significant health concern that affects a large population worldwide. Although imaging techniques such as magnetic resonance and endoscopic ultrasound are commonly used for gallstone detection, they can be expensive and sometimes yield inaccurate results. These techniques also require expensive equipment and specialized expertise, making them less accessible. In this study, we introduce a new method for predicting gallstones from bioimpedance and laboratory data using artificial intelligence instead of imaging techniques. Our dataset includes 38 parameters, including age, gender, some blood parameters, and bioimpedance data. The selection of differentiable features for identifying gallstone disease is a critical step in enhancing classification performance. In this study, we propose a new ensemble feature selection method based on metahuristic and statistical methods. To achieve this, we first apply four statistical feature selection methods, which are minimum redundancy maximum relevance, chi-square test, analysis of variance, the Kruskal-Wallis, followed by four metaheuristic feature selection methods, which are grey wolf optimization, whale optimization algorithm, harris hawk optimization, and particle swarm optimization. The proposed ensemble feature selection method selects features that are chosen by at least seven methods among four statistics and four metaheuristic methods to determine the most effective features for identifying gallstones. To evaluate the effectiveness of the selected features, we apply logistic regression, Naïve Bayes, support vector machine, decision tree, k-nearest neighbor, gradient boosting, Adaboost, random forest, and multilayer perceptron methods for classification. The random forest using the proposed ensemble feature selection method achieves the highest accuracy of 90.62% by using 11 features instead of 38. The proposed feature selection method significantly reduces classification complexity and improves overall performance of the machine learning methods. The superiority of the proposed method is also highlighted by comparing it with state-of-the-art studies that predict gallstones based on precision and number of features.

Keywords: Gallstone Disease, Machine Learning Algorithms, Meta-heuristic Algorithms, Feature Selection, Clinical Decision Support Systems.

METASEZGİSEL MAKİNE ÖĞRENME ALGORİTMALARI İLE SAFRA TAŞI TESPİTİ

ÖZ

Safra taşı hastalığı, dünya çapında büyük bir nüfusu etkileyen önemli bir sağlık sorunudur. Görüntüleme teknikleri safra taşı tespiti için yaygın olarak kullanılsa da pahalı olabilmekte ve bazen yanlış sonuçlar verebilmektedir. Etkinlikleri operatör becerisi, görüntüleme çözünürlüğü, taş bileşimi ve hastaya özgü koşullar gibi faktörlere bağlıdır. Manyetik rezonans görüntüleme ve endoskopik ultrason gibi teknikler de pahalı ekipman ve özel uzmanlık gerektirdiğinden daha az erişilebilirdir. Bu çalışmada, görüntüleme teknikleri yerine yapay zeka kullanarak biyoimpedans ve laboratuvar verilerinden safra kesesi taşlarını tahmin etmek için yeni bir yöntem sunuyoruz. Veri setimiz yaş, cinsiyet, kan parametreleri ve biyoimpedans verileri dahil olmak üzere 38 parametre içermektedir. Safra taşı hastalığını tanımlamak için ayırt edilebilir özelliklerin seçilmesi, sınıflandırma performansını artırmada kritik bir adımdır. Bu çalışmada, yeni bir topluluk özellik seçimi yöntemi öneriyoruz. Bunu başarmak için, öncelikle minimum redundansi maksimum alaka, ki-kare testi, varyans analizi, Kruskal-Wallis olmak üzere dört istatistiksel özellik seçim yöntemi ve ardından gri kurt optimizasyonu, balina optimizasyon algoritması, harris hawk optimizasyonu ve parçacık sürüsü optimizasyonu olmak üzere dört metasezgisel özellik seçim yöntemi uyguluyoruz. Önerilen topluluk özellik seçim yöntemi, safra kesesi taşlarını tanımlamak için en etkili özellikleri belirlemek üzere dört istatistik ve dört metasezgisel yöntem arasından en az yedi yöntemle seçilen özellikleri seçer. Seçilen özelliklerin etkinliğini değerlendirmek için lojistik regresyon, Naive Bayes, destek vektör makinesi, karar ağacı, k-en yakın komşu, gradyan artırma, Adaboost, rastgele orman ve sınıflandırma için çok katmanlı perceptron yöntemleri uygulanmaktadır. Önerilen topluluk özellik seçimi yöntemini kullanan rastgele orman, 38 özellik yerine 11 özellik kullanarak 90.62% ile en yüksek doğruluğa ulaşmaktadır. Önerilen özellik seçim yöntemi, sınıflandırma karmaşıklığını azaltmakta ve doğruluk ile seçilen özellik sayısı açısından mevcut yöntemlere göre üstünlük sağlamaktadır.

Anahtar kelimeler: Safra Taşı Hastalığı, Makine Öğrenmesi Algoritmaları, Meta-sezgisel Algoritmalar, Özellik Seçimi, Klinik Karar Destek Sistemleri.

CONTENTS

M.Sc. THESIS EXAMINATION RESULT FORM ii
ETHICAL DECLARATION iii
ACKNOWLEDGEMENTS iv
ABSTRACT v
ÖZ vi
NOMENCLATURE x
LIST OF TABLES xii
LIST OF FIGURES xiii

CHAPTER 1 – INTRODUCTION 1

1.1 Motivation 1
1.2 Problem Statement 2
1.3 Contributions 2
1.4 Organization of the Thesis 3

CHAPTER 2 – LITERATURE REVIEW 4

CHAPTER 3 – MATERIALS AND METHODS 11

3.1 Dataset and Features 11
3.2 Proposed Ensemble Feature Selection Method 13
3.2.1 Statistical Methods for Feature Selection 16
3.2.1.1 Minimum Redundancy Maximum Relevance (MRMR) 16
3.2.1.2 Chi-square (Chi^2) 18
3.2.1.3 Analysis of Variance (ANOVA) 20
3.2.1.4 Kruskal-Wallis 22
3.2.2 Metaheuristic Algorithms for Feature Selection 24
3.2.2.1 Grey Wolf Optimization (GWO) 24
3.2.2.2 Harris Hawk Optimization (HHO) 26
3.2.2.3 Particle Swarm Optimization (PSO) 29
3.2.2.4 Whale Optimization (WO) 30

3.3 Machine Learning Models 34
3.3.1 Machine Learning Algorithms 35
3.3.1.1 Logistic Regression Algorithm 36
3.3.1.2 Naive Bayes Algorithm 38
3.3.1.3 Support Vector Machine Algorithm 40
3.3.1.4 Decision Tree Algorithm 42
3.3.1.5 K-Nearest Neighbor Algorithm 44
3.3.1.6 Gradient Boosting Algorithm 46
3.3.1.7 Adaboost Algorithm 47
3.3.1.8 Random Forest Algorithm 49
3.3.1.9 Multilayer Perceptron Algorithm 51

CHAPTER 4 – EXPERIMENTAL RESULTS 54

4.1 Results of the machine learning classifiers without feature selection 54
4.2 Results of the machine learning classifiers using statistical feature selection methods 56
4.3 Results of the machine learning classifiers using metaheuristic feature selection methods 59
4.4 Results of Proposed Ensemble Feature Selection Method 63
4.5 Comparison with the-state-of-the-art gallstone detection methods 65

CHAPTER 5 – CONCLUSION AND FUTURE STUDIES 68

5.1 Conclusion Remarks 68
5.2 Future Work 70
5.2.1 Dataset Generalization and Population Diversity 70
5.2.2 Incorporation of Multimodal and Longitudinal Data 71
5.2.3 Data Quality, Missing Values, and Class Imbalance Handling 71
5.2.4 Enhanced Feature Engineering and Metaheuristic Exploration 72
5.2.5 Advancement of Machine Learning Models and Optimization Techniques 73
5.2.6 Interpretability and Clinical Trust via Explainable AI 73

5.2.7 Deployment in Real-Time and Embedded Systems 74
5.2.8 Integration into Clinical Decision Support Systems (CDSS) 74
5.2.9 Clinical Trials, Usability Testing, and Physician Engagement 75
5.2.10 Patient-Facing Applications and Remote Healthcare 75
5.2.11 Model Maintenance, Updating, and Drift Detection 76

REFERENCES 77

APPENDICES 91

Appendix A – Features of the Dataset 92
Appendix B – List of library and module versions used 93

CURRICULUM VITAE 94

NOMENCLATURE

Greek Letter Symbols

α	Learning rate used in model training
β	Parameter represents coefficients or weights in machine learning models.
γ	In the SVM algorithm, the parameter determines the effect of a sample on the decision boundary.

Acronyms

Acc	Accuracy
AI	Artificial Intelligence
ALP	Alkaline Phosphatase
ALT	Alanine Aminotransferase
ANOVA	Analysis of Variance
AST	Aspartate Aminotransferase
BMI	Body Mass Index
CDSS	Clinical Decision Support Systems
Chi^2	Chi-square Test
CRP	C-reactive Protein
CT	Computed Tomography
DT	Decision Tree
F1	F1 Score
FN	False Negative
FP	False Positive
GB	Gradient Boosting
GWO	Grey Wolf Optimization
HDL	High-Density Lipoprotein
HHO	Harris Hawk Optimization
KNN	K-Nearest Neighbor
LDL	Low-Density Lipoprotein
LR	Logistic Regression
MLP	Multilayer Perceptron
MRI	Magnetic Resonance Imaging

MRMR	Minimum Redundancy Maximum Relevance
NB	Naive Bayes
P	Precision
PSO	Particle Swarm Optimization
R	Recall
RF	Random Forest
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
WOA	Whale Optimization Algorithm

LIST OF TABLES

Table 4.1 Performance metrics of the machine learning methods 54

Table 4.2 Performance of the machine learning classifiers without feature selection 55

Table 4.3 Results of the machine learning methods when statistical feature selection methods are used. 56

Table 4.4 Results of the machine learning methods when the metaheuristic algorithms are used in the feature selection process 60

Table 4.5 Results of the machine learning methods when the proposed feature selection method is used. 65

Table A.1 Features of the Dataset 92

Table B.1 List of library and module versions used 93

LIST OF FIGURES

Figure 3.1 Fundamental steps of the proposed method for identifying patients with gallstones 12

Figure 3.2 Proposed feature selection method for detecting the optimum set of features identifying patients with gallstones 14

Figure 4.1 Confusion matrices for each machine learning classifier without feature selection. 55

Figure 4.2 Confusion matrices for each machine learning classifier when MRMR is used. 57

Figure 4.3 Confusion matrices for each machine learning classifier when Chi^2 is used. 58

Figure 4.4 Confusion matrices for each machine learning classifier when ANOVA is used. 58

Figure 4.5 Confusion matrices for each machine learning classifier when Kruskal-Wallis is used. 59

Figure 4.6 Accuracy results of the machine learning methods using metaheuristic feature selection methods 60

Figure 4.7 Confusion matrices for each machine learning classifier when GWO is used. 61

Figure 4.8 Confusion matrices for each machine learning classifier when PSO is used. 62

Figure 4.9 Confusion matrices for each machine learning classifier when WOA is used. 62

Figure 4.10 Confusion matrices for each machine learning classifier when HHO is used. 63

Figure 4.11 Confusion matrices for each machine learning classifier when the proposed ensemble feature selection method is used. 64

Figure 4.12 Accuracy results of the machine learning methods when all features (without feature selection) are used, and ANOVA, WOA as well as the proposed ensemble feature selection method are applied in the feature selection process 66

CHAPTER 1

INTRODUCTION

1.1 Motivation

Gallstones are solid crystalline substances that form due to the accumulation of gallbladder components such as cholesterol or bilirubin. Although gallstones can be treated relatively easily, the most common approach involves gallbladder removal. This can lead to severe proximal abdominal pain, especially on the right side, which greatly worsens the quality of life for the patient. Moreover, a delayed diagnosis could complicate treatment, making it harder to perform surgery. These elements reinforce the importance of early diagnosis as a critical aspect in the successful treatment of the disease.

With global population increasing, the number of doctors per capita has steadily declined. This has resulted in heavier workloads for medical professionals, increasing the risk of misdiagnosis, incorrect treatments, and even patient fatalities. Research in artificial intelligence (AI) based medical diagnostics has therefore gained considerable attention due to the high cost of diagnostic tests and the time consuming nature of conventional methods. Artificial intelligence techniques provide highly accurate solutions for various diseases in the healthcare field. Several studies utilizing machine learning and AI technologies have been published to assist healthcare professionals in interpreting data and predicting outcomes [1]. In recent years, AI has been effectively used for classifying malignant skin tumors [2], diagnosing various tumors using imaging data [3], detecting retina-related diseases [4], and diagnosing gynecological cancers [5]. In the case of gallstones, Artificial intelligence based approaches can study large volumes of healthcare data, find out early symptoms, and help physicians to make accurate decisions. In the long run, this improves patient outcomes and decreases the risk of clinical errors. Many diseases, including gallstone disease, can be diagnosed based on symptom patterns using AI methods.

1.2 Problem Statement

Gallstone disease throughout the world is a common health issue and if left untreated can lead to complicated serious conditions. This disease, which results from a crystallization of cholesterol or bilirubin in the gallbladder, can obstruct the bile ducts, lead to infections and pancreatitis. Early diagnosis allows the treatment process to be carried out before the disease becomes more severe and is useful to minimize unnecessary surgical procedures.

Imaging studies including ultrasonography and computed tomography are commonly used to diagnose the disease. But they are expensive, demanding special equipment and highly skilled specialists, and are not widely available in all health facilities. In particular, diagnostic delays deny timely treatment and generate extra costs to the health system in resource-constrained regions. Furthermore, the operator dependency of the diagnostic process causes variations in interpretations among experts, leading to reduced clinical accuracy. These gaps hinder patients' timely access to treatment through long waiting times in the diagnostic process, reducing the effective use of health services. This problem is compounded in low-resource and mid-resource settings, and the proposed method efficiently detects patients with gallstones.

1.3 Contributions

We propose a machine learning based gallstone detection method which is an alternative to ultrasound or other visual imaging techniques and can be done based on bioimpedance and laboratory data in this paper. The contributions of this research can be summarized as:

- The selection of differentiable features for identifying gallstones is a critical step in enhancing classification performance. In this study, we propose a new ensemble feature selection method. To achieve this, we first apply four statistical feature selection methods, followed by four metaheuristic feature selection methods. We

select features that are chosen by at least seven methods to determine the most effective features for identifying gallstones.

- We propose a robust gallstone prediction method as an alternative to imaging techniques, using machine learning techniques enhanced by metaheuristic feature selection methods
- The proposed method efficiently identifies patients with gallstones by looking at 11 features, and the superiority of the proposed method is also highlighted by comparing with the state-of-the-art studies detecting gallstones

1.4 Organization of the Thesis

This thesis is organized into five chapters which together provide an integrated view of gallstone detection using metaheuristic-assisted machine learning methods. Chapter 1 introduces articulating motivation, defining the research problem, and outlining the primary contributions. The proposed classification strategy for identifying patients with gallstones is detailed in Chapter 3. Chapter 4 describes the experiment results, assesses the efficacy of the proposed approach, and reviews the overall performance of the existing techniques. Chapter 5 closes with concluding remarks and identifies avenues for future research.

CHAPTER 2

LITERATURE REVIEW

In this chapter, we briefly overview studies detecting gallstones using artificial intelligence based approaches.

Salem et al. [6] performed a metabonomics study to identify certain metabolites and clinical characteristics related to gallstone disease. They used deep machine learning models to predict the risk of developing gallstones. They also examined the influence of gender as a contributing factor for the gallstone formation. According to the findings of their experimental study, their proposed model based on convolutional neural network (CNNs) achieved 83% accuracy in gallstone prediction. In another study, Yao et al. [7] applied deep learning techniques to analyze clinical textual data for identifying gallstone disease. By enhanching word embeddings and medical ontologies, their proposed CNN model successfully classified gallstone-related cases with an F1 score of 98.37%. They were also emphasized the importance of the linguistic features and domain-specific knowledge to improve classification performance, especially in gallbladder-related cases. Liew et al. [8] compared artificial neural networks (ANN) to logistic regression (LR) for predicting gallbladder disease among obese patients. Their study used the data of 117 patients who underwent bariatric surgery between 1999 and 2005, 12 of whom were diagnosed with gallstone disease. Due to this model performance evaluation, the ANN model obtained the most accuracy of 97.14%, when the LR model could only achieve an accuracy of 88.2%. Their results showed that the ANN model was significantly more effective than the other models in predicting gallstone disease. Using arbitrary cut points for each level of risk, the authors found that significant predictive factors included diastolic blood pressure, chronic inflammation, and HbA1c levels. Furthermore, the study highlighted that ANN models outperform traditional statistical methods, reinforcing their potential as a robust tool for gallstone prediction. Obaid et al. [9] explored the performance of a set of deep learning models for gallbladder disease classification with ultrasound images. They used 10,692 ultrasound images in

the dataset collected from three hospitals and two clinical centers. In addition, the data included a total of nine different gallbladder diseases, including gallstones. Following that, image processing techniques and ROI identification processes were applied and the trained Visual Geometry Group with 16 layers, Inception Version 3, Residual Network with 152 layers, and MobileNet models using transfer learning techniques. Their results showed that the model MobileNet classifies images with the most success rate of 98.35% accuracy. Deep learning based image analysis techniques potentially are viewed by them as source of high accurate diagnostic help for gallstone disease in the context of clinical decision support system. Obaid et al. [10] performed a comparison analysis between machine learning and deep learning techniques respectively in diagnosis of gallbladder diseases. As with others studies in scope of the study, 44 different research studies have been examined, and the accuracy rates of the methods have been evaluated. The datasets used included medical imaging data such as ultrasound, but also computed tomography (CT), magnetic resonance imaging (MRI), and health records. One of the main pathologies studied was gallstone disease, emphasizing the importance of AI-based approaches for both early diagnosis and disease management. According to the study, the machine learning techniques achieved an accuracy rate of between 69.67% and 96.67%, while the maximum accuracy for the deployment of deep learning models was 98.77%. Showing that AI-based diagnostic methods represent a viable alternative to traditional methodologies, these insights suggest a distinct advantage of such systems, namely as applied to gallstone disease.

Thomas et al. [11] performed a machine learning model to provide decision support in the management of patients with gallstone disease. They utilized identification of critical risk factors extracted from the electronic health records (EHR) to predict disease progression and the necessity for endoscopic retrograde cholangiopancreatography (ERCP). They introduced a new neural network structure known as ModCNN (Modified Cascade Neural Network) and reported so far not ANN (Artificial Neural Networks) and CCNN (Cascade-Correlation Neural Networks) comparative results. The accuracies of ModCNN (96.42%) were the best among ANN (89.65%) and CCNN (93.24%); thus, experimental results confirmed that the

ModCNN outperformed ANN and CCNN. The research focused on how measures that predict patient risk can be integrated with computerized decision support systems (CDSS), which directs attention to diagnostic and therapeutic options that might be overlooked leading to reduce medical errors and improve healthcare. Blum et al. [12] evaluated the diagnostic efficacy of machine learning models (MLMs) in detecting choledocholithiasis; and compared them with the American Society for Gastrointestinal Endoscopy (ASGE) guidelines. Although model accuracies varied from 19% to 97% in the review, studies with low bias described maximum accuracies between 70% and 85%. The MLMs were mostly compliant with the ASGE criteria, especially age, total bilirubin, and common bile duct size as predictors. However, this study also points to the importance of greater methodological transparency and validation in future MLM applications. Vukicevic et al. [13] introduced an automated expert system of building artificial neural networks (ANNs) aimed at predicting common bile duct stones (CBDS) during gallstone surgery. This method utilized a genetic algorithm for ANN configuration and backward feature selection. Their ANN model with data from 303 patients successfully identifies total bilirubin, CBD diameter, and acute cholecystitis as significant predictors. The resulting ANN was determined to have higher positive predictive power and clinical viability compared to a logistic regression and decision tree models, adding to the growing body of knowledge regarding the benefit of automating feature optimization for clinical AI methods. Dalai et al. [14] used generation of random forest-based models for predicting the presence of choledocholithiasis (via ERCP) using non-invasive imaging data, including ultrasound, CT, and MRCP. Demographic, biochemical, and radiological features were used in their models, resulting in total bilirubin levels and common bile duct (CBD) diameter being prominent predictors. Overall, the highest sensitivity and specificity were achieved with the random forest model (77.3% and 75.0%, respectively), which exceeded the performance of the ASGE high-likelihood criteria (90.9% sensitivity) but at a cost of significantly lower specificity (25.0%). These results indicated a better balance of machine learning diagnostic performance, highlighting its potential as a non-invasive alternative for choledocholithiasis risk stratification as set out in current guidelines. Sachs et al. [15] used machine learning for the prediction of pediatric choledocholithiasis in 10 pediatric-based institutions.

One such model achieved an AUC of 0.935 using an Extra-Trees classifier trained on a 9-feature model that included clinical variables such as AST, ALT, lipase, and CBD measurements. Their study emphasized an increasing demand for tailored, data-driven tools in pediatric diagnostics, especially considering the rising prevalence of childhood obesity, an important risk factor for gallstone diseases.

Kocer et al. [16] investigated the effectiveness of the You Only Look Once (YOLO) Version 5 algorithm in identifying various diseases, including gallstones, from abdominal computed tomography (CT) images. In their study, CT images obtained from 1,200 patients were utilized to detect 11 different conditions, namely gallstones, acute cholecystitis, kidney stones, and appendicitis. Their proposed YOLO algorithm achieved an accuracy rate of 97.80% for gallstone detection, outperforming previous-generation object detection models. The model enhanced object recognition precision by leveraging anchor box structures and Non-Maximum Suppression (NMS) techniques. Their proposed method validates a promising algorithm to be used as a robust and early detection tool for gallstone disease, in their case, using the YOLO Version 5 algorithm. Yu et al. [17] introduced lightweight deep neural networks to help in ultrasound diagnostic of cholelithiasis and cholecystitis. They used datasets containing 89,341 ultrasound images from Taipei Medical University Hospital and 85,870 ultrasound images for gallstone diagnosis by using the Single Shot MultiBox Detector-Feature Pyramid Network-Residual Network that has 50 layers (SSD-FPN-ResNet-152) model and cholecystitis diagnosis by using the MobileNet model. The MobileNet model they proposed achieved high performance, including detection of cholecystitis with an area under the curve (AUC) of 94% and a sensitivity of 88%. The SSD-FPN-ResNet-50 model performed similarly, successfully classifying gallstone cases with a sensitivity of 92%, and a specificity of 80% These results demonstrate that deep learning based lightweight models have their great potential to contribute to the interpretation of ultrasound images especially within emergency medical scenarios. Himi et al. [18] developed a system capable of predicting common diseases using data collected from smartwatch sensors. Their study sought to predict 12 diseases, among them gallstones, through biometric data like heart rate, body temperature, and blood oxygen saturation. Performance of the

trained machine learning models showed that the RF algorithm performed the best with an accuracy of 99.4%, demonstrating the utility of their dataset towards effective prediction of gallstone disease. These findings emphasized the potential of wearable health technologies in optimizing disease prediction and their integration into clinical decision support systems. Zi et al. [19] aimed to detect and classify gallstone disease using an AI-based system. In their study, MobileNet Version 2, a lightweight deep learning model designed for mobile and embedded devices, and Single Shot MultiBox Detector Lite, an object detection algorithm capable of identifying multiple objects at different scales in a single pass, were utilized for the classification of gallstone types. The CT images belonging to 673 patients were analyzed, and these models successfully identified different stone types. Yao et al. [20] trained a CNN (Convolutional Neural Network) model to predict the chemical composition of gallstones using imaging data from medical IoT (Internet of Things) applications. This includes identification of the direct impact on treatment decisions of cholesterol, bile pigment, mixed, and rare stone type (for stones of unexplained composition). The model used ultrasound, CT, and MRI images of gallstones and learned distinguishing characteristics through a supervised deep learning approach. This study highlighted that the chemical composition of gallstones, which could previously only be determined after surgical extraction, could instead be predicted non-invasively, using energy dispersive X-ray spectra extracted such as smart imaging data and then applied to a convolutional neural network model, with implications for an increasingly individualized approach to gallstone treatment in the future. Zhou et al. [21] described a rapid diagnostic engine for choledocholithiasis and recurrence prediction by analysis of raw bile fuelled with SERS (Surface-Enhanced Raman Spectroscopy through machine learning. Using SERS spectra of 920 spectra across 46 patients, PCA and SVM were used to classify three types of stones with a diagnostic accuracy of 99.2% and a recurrence prediction accuracy of 95.2%. By relying on Pearson correlation they investigated inter-component interactions to derive insight helping in establishing better disease mechanisms. Turki et al. [22] introduced a large-scale ultrasound image dataset named UIdataGB and proposed an elaborate deep learning-based framework for gallbladder disease classification. The dataset contains 10,692 high-quality images from 1,782 patients acquired from four different ultrasound devices and nine distinct

gallbladder disease classes. The images were preprocessed, normalized, and augmented for the purposes of optimal machine learning performance. Based on their dataset, multiple CNN (Convolutional Neural Network) architectures were employed, which aimed to perform automatic classification in various types of diseases. Using a low-image dataset, the proposed system achieved high diagnostic accuracy and reliability to assist clinical decision-making by giving computer-aided opinions on gallbladder diseases with gallstones, cholecystitis, and carcinoma. Vanitha et al. [23] utilized the SWFCM (Spatially Weighted Fuzzy C-Means Clustering) algorithm along with Support Vector Machines (SVM) for the segmentation and classification of gallstones in ultrasound images. The images were a total of 100 ultrasound, in their study, with the best accuracy of 88.56% with the SVM classifier. The results showed that in terms of segmentation, the SWFCM algorithm surpassed other classical methods. Their study holds significant potential to enhance clinical diagnostic processes by providing an automated and cost-effective approach for detecting gallstones in ultrasound images. Yanto et al. [24] developed an unsupervised multi-stage segmentation algorithm to detect gallstone disease in magnetic resonance cholangiopancreatography (MRCP) images. The algorithm employs K-Means clustering and morphological techniques to accurately segment gallstones within MRCP images. Segmentation performance was evaluated using object separation accuracy, achieving an accuracy rate of 90%. Reports of their findings showed the proposed method improved accuracy in the prediction of diagnostic processing as well as reduced overall time of the manual segmentation process which can assist and facilitate clinical studies. Song et al. [25] proposed a new deep learning-based architecture U-NeXt which is used to segment gallstones from CT images. The U-NeXt model employs Dense Skip Connections (DSC), Attention Up-Sampling Block, and Skip Spatial Pyramid Pooling (SkipSPP) module to retrieve multiscale features with the goal of achieving better segmentation performance. Their study utilized 5350 CT images sourced from 726 patients, with gallstone annotations co-labelled by three experienced health experts. According to the results of experiments carried out, the intersection over union (IoU) improved by 7.03% when comparing U-NeXt to the baseline U-Net model, while the performance when comparing to traditional segmentation methods increased by up to 24%. Furthermore,

they built and released a new annotated gallstone CT dataset, and showed that our new network, U-NeXt, is capable of accurately segmenting both typical and atypical gallstone structures. Their study findings suggested the proposed proximity model can provide robust and automated segmented outcomes and thus enable reliable clinical diagnosis and improved efficiency in gallstone disease detection.

Finally, Esen et al. [26] was proposed a study for gallstone prediction through machine learning approaches, based on laboratory and bioimpedance data. Their dataset comprised of 319 patients with suspected gallstone disease; 161 patients were diagnosed with gallstones. Their dataset included 38 features such as height, weight, age, bioimpedance data, and blood test results. Their experimental results indicated that the accuracy of the predictive models usage to predict gallstones using the gradient boosting method was 85.42%. On the other hand, in this thesis, we use the same dataset as in the study of Esen et al. [26], and suggest the use of metaheuristic techniques to choose the best parameters to detect gallstones from the 38 features. Our proposed method identifies gallstones using fewer features compared to Esen et al. [26], while demonstrating improved accuracy and effectiveness in predicting gallstone disease.

CHAPTER 3

MATERIALS AND METHODS

In this chapter, we provide information about the dataset used in this study and the proposed method for gallstone detection. The proposed method comprises two main components: feature selection and classification. Feature selection [27] is the process of identifying the most relevant and informative features for detecting patients with gallstones in the dataset. This process aims to eliminate redundant or irrelevant features, hence enhancing the classification model's performance. Additionally, feature selection improves computational efficiency by reducing the dimensionality of the data, leading to faster model execution. Furthermore, achieving high performance with a reduced number of features enhances the interpretability of the classifier model and facilitates a more comprehensive understanding of the classification process. After determining the optimal features for predicting gallstone presence in patients, the second step involves classifying the patients based on the presence of gallstones using machine learning techniques. The main steps of the proposed method are presented in Figure 3.1. In the following subsections, we explain these steps in detail.

3.1 Dataset and Features

The dataset used in this study is Gallstone Dataset, obtained from the website of The UCI Machine Learning Repository at https://archive.ics.uci.edu/dataset/1150/gallstone-1. Data of participants were retrieved from the Medical Records of 319 patients who had applied to the Internal Medicine Outpatient Clinic of Ankara VM Medical Park Hospital between June 2022 and June 2023. The study was approved by the Ankara City Hospital Medical Ethics Committee (E2-23-4632) abstract, and was conducted in accordance with rules ethical. Anonymized data were extracted from the hospital systems and written informed consent was obtained from all participants. Specialist physicians confirmed the diagnoses in patients with suspected gallstone disease.

Based on Traditional Biochemical Tests and Bioimpedance Measurements, the

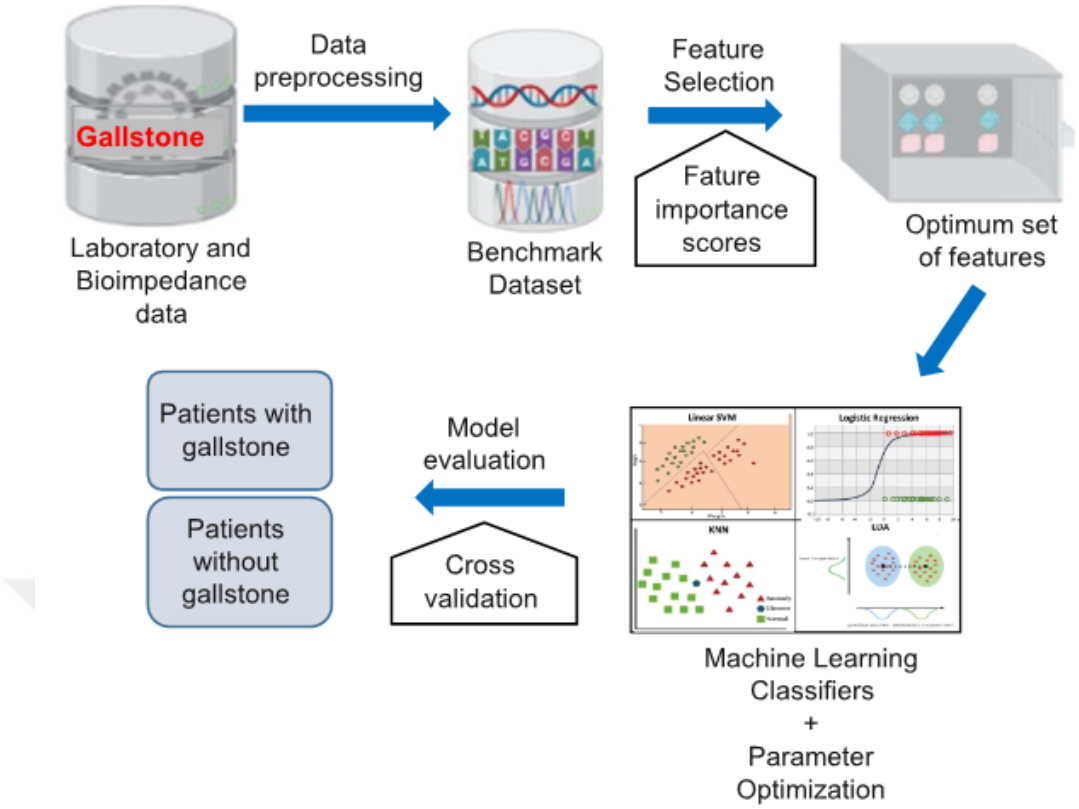

Figure 3.1 Fundamental steps of the proposed method for identifying patients with gallstones

Dataset Contains 38 Features (number of days of hospitalization, age, sex, and lifestyle among them). Compared to previous studies where only biochemical parameters were typically used, both biochemical and bioimpedance measurements were considered here to re-evaluate the diagnosis process more holistically. By this we use this as the starting point to clarify the diagnostic mediating effects surrounding margin composition in gallstone disease.

In this study, we analyzed demographic variables, chronic disease history, body composition metrics obtained through bioimpedance analysis, and biochemical parameters from laboratory tests. It is clear from the description that this is a words problem hence we create the dataset accordingly and use the presence or absence of gallstones as the outcome variables here as it is a binary classification. This dataset, was used to discover potential clinical biomarkers for the early diagnosis of gallstone disease, through machine learning-based predictive modeling for gallstone disease detection. Demographic variables include height, weight, age, gender, and BMI (Body Mass Index). The history of chronic diseases covers conditions such as

hypothyroidism, hyperlipidemia, diabetes mellitus, and coronary artery disease. Variables derived from bioimpedance analysis assess body composition, including total body water, extracellular water, intracellular water, muscle mass, protein content, total body fat percentage, visceral fat area, and hepatic fat accumulation. Laboratory test results encompass glucose, total cholesterol, LDL (Low-Density Lipoprotein), HDL (High-Density Lipoprotein), triglycerides, AST (Aspartate Aminotransferase), ALT (Alanine Aminotransferase), ALP (Alkaline Phosphatase), creatinine, glomerular filtration rate, CRP (C-reactive Protein), hemoglobin, and vitamin D levels.

Since the dataset had no missing data then no data completion method was applied and, due to the balanced distribution of sick and healthy individuals, there was no additional data balancing step. Furthermore, transformations (such as scaling and normalization) were not applied and analyses were conducted on raw data.

The inclusion of bioimpedance measurements within the dataset is a well-defined strength of the dataset in terms of the opportunity to investigate, and reduce the risk of associations between gallstone disease and biometric measurements; low error rates during preprocessing owing to the complete absence of missing data; and sufficiently sized datasets for application of machine learning models. However, data in this study was only collected in a specific hospital, thus can not be generalized. In addition, the model has not yet been evaluated in diverse populations, as it has not yet been compared with an external dataset.

3.2 Proposed Ensemble Feature Selection Method

In this study, we perform four statistical feature selection methods: MRMR (Minimum Redundancy Maximum Relevance), Chi^2 (Chi-square), ANOVA (Analysis of Variance), and Kruskal-Wallis. Furthermore, we apply four state-of-the-art metaheuristic methods: GWO (Grey Wolf Optimization), HHO (Harris Hawk Optimization), WO (Whale Optimization), and PSO (Particle Swarm Optimization).

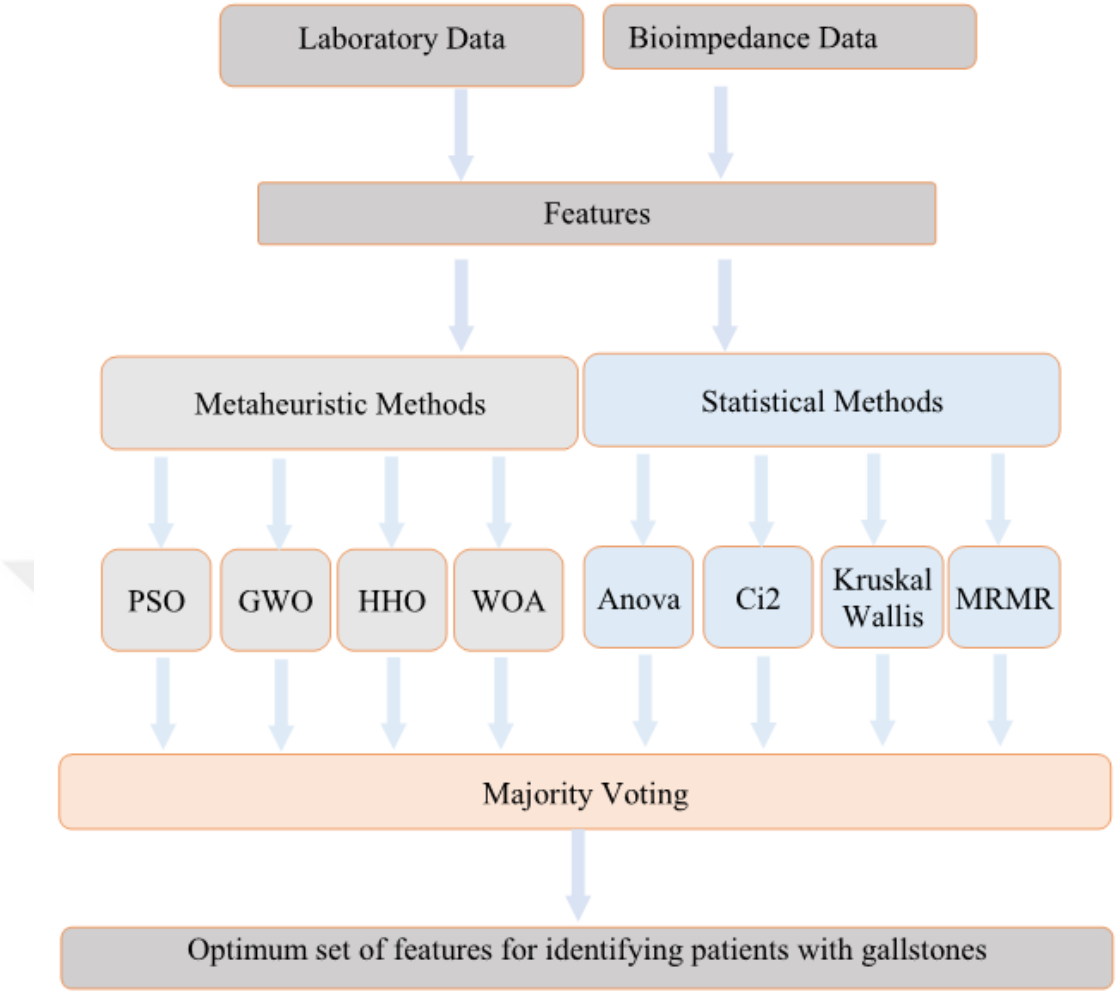

Figure 3.2 Proposed feature selection method for detecting the optimum set of features identifying patients with gallstones

Next, we choose the features that are selected at least 7 feature selection methods among four statistics and four metaheuristics methods. The key procedures of the proposed feature selection method are illustrated in Figure 3.2. Statistical feature selection methods only worked to find important features by analyzing the coefficient of features with the target variable. But traditional approaches often consider variables as independent, generally do not adequately account for variable interactions. Thus, besides traditional statistical methods, meta-heuristic algorithms have been applying to minimize the selection of the features. The complexity of these search spaces to find optimal solutions in the search space improved the performance of the model with the help of these algorithms.

In the case of large data sets, metaheuristic algorithms can add stronger characteristics,

and can discover some complex interrelations among the variables, which cannot be achieved through classical statistical techniques. This multi-faceted analysis and optimization of features optimize model performance with the set configuration of hyperparameters, adaptable to various data shapes. This means that, for example, the GWO algorithm contributes to improved model accuracy by providing a more balanced approach to the selection of variables, while the HHO is superior in speed, especially within large search spaces. Since there are less variables on data sets, the PSO algorithm is capable of eliminating irrelevant features through an efficient selection process, while WOA shows flexible and adaptive solutions in various structures that enhance the model performance. As such, the benefits of these algorithms allow more reliable and computationally efficient diagnostics processes by means of permitting access to the identification of the most significant variables, as it is natural to expect in high-dimensional data sets.

Dokeroglu et al. [28] published a review on the application of metaheuristic algorithms to feature selection. In summary, they suggest that the use of metaheuristics in feature selection will continue to expand rapidly in the coming years, ultimately leading to the development of methods that are considered state-of-the-art. For Feature selection, we perform four state-of-the-art metaheuristic methods in this study.

Lastly, four statistical and four metaheuristics methods for feature selection were used, in total eight different solutions. Combining the two approaches, and thus avoiding using only a statistical or only a metaheuristic method generated the most successful outcomes. In this way, the model's accuracy was significantly improved with the use of features chosen by the majority of eight different methods. It also indicates that instead of focusing on feature selection using one method we should be integrating both the firm groundwork laid out around statistical analysis and the immense investigative potential associated with metaheuristic-based methods, particularly in high-dimensional data. By considering the interplay between variables in a holistic manner, this approach can facilitate more meaningful assessments of how components in a system interact to influence overall behavior, leading to a more integrated model that enhances the reliability and efficiency of diagnostic systems.

3.2.1 Statistical Methods for Feature Selection

In this subsection, we explain statistical feature selection methods that are minimum redundancy maximum relevance, Chi-square, analysis of variance, and Kruskall-Wallis performed in this thesis.

3.2.1.1 Minimum Redundancy Maximum Relevance (MRMR)

MRMR [29] is one of the most popular feature selection approaches in the high dimensional setting. This method aims to reduce the redundancy of the data, so that if a feature is present, the information it contains is not too redundant with the value of selected features while also maintaining the most fitted features to the target variable. It uses the Mutual Information metric, and follows a forward-strategy approach for feature selection. The first step picks the feature which has the greatest dependency to the target variable, in steps adds new features at each time point, and ranks them mutually with the ones already chosen. This approach is especially on classification problems where it seeks to select features those deliver the most meaningful contribution due to their relevant to the class.

MRMR is advantageous as it helps not only with the features that will help increase the model accuracy as well as balance it by helping with more efficient and representative features, but also helps with computational efficiency of the model, especially when working with high-dimensional data, breaking down the computational cost. This leads to a reduction of multicollinearity and the generation of more stable and generalizable models since we stop using very similar features. Overall it helps to prevent the algorithm from overfitting. While MRMR can be computationally intensive in larger datasets owing to the exhaustive calculations of inter-variable dependence, the method lays the groundwork for data preprocessing and ensures that the analysis is unaffected by noisy or irregular features. It is widely used in domains where high-dimensional data is normal, including bioinformatics, image-processing and even text-mining. In Algorithm 1, pseudocode summarizing the steps of the MRMR used in this study is presented.

Furthermore, a study of the methods found in the literature to detect for example diseases such as thymoma [30], lung cancer [31], and atrial fibrillation [32], reveals that the characteristic selection have a direct impact on the success of the model. According to these studies, the techniques applied for selecting the top related features from complex and high-dimensional biomedical data greatly increase the accuracy of the classification method and also significantly decrease the model training time. As an illustration, while working with Transformer-based models for detecting thymoma, the complexity of input was reduced to obtain better model generalization capabilities. Likewise, experiments with CNN models and, in the present case chest CT images for lung cancer detection, found that accuracy rates were improved if the number of features offered to the classifier was decreased from the original input of all training data to a second offering of fewer attributes, thereby increasing interpretability by reducing complexity. Sensible feature elimination based on heartbeat data has also led to the development of lightweight models that detect the disease effectively while contributing to early diagnosis in research studies on the detection of atrial fibrillation. These examples illustrate that there are multiple metrics, not only for accuracy, but also related to computational burden, model simplicity, and clinical applicability, associated with these gains.

Algorithm 1 MRMR Method [29]

Input: Feature set $X = \{x_1, x_2, ..., x_n\}$, target variable Y, number of selected features k
Output: Selected feature subset S
Initialize $S \leftarrow \emptyset$, $R \leftarrow X$
Compute mutual information $I(x_i, Y)$ for all $x_i \in X$
Select $x^* = \arg\max_{x_i \in R} I(x_i, Y)$
Add x^* to S and remove it from R
while $|S| < k$ **do**
    for each $x_j \in R$ **do**
        Compute relevance: $\text{Relevancy}(x_j) = I(x_j, Y)$
        Compute redundancy: $\text{Redundancy}(x_j) = \frac{1}{|S|} \sum_{x_s \in S} I(x_j, x_s)$
        Compute MRMR score: $\text{MRMR}(x_j) = \text{Relevancy}(x_j) - \text{Redundancy}(x_j)$
    end for
    Select $x^* = \arg\max_{x_j \in R} \text{MRMR}(x_j)$
    Add x^* to S and remove it from R
end while
Return S

3.2.1.2 Chi-square (Chi²)

Chi² test [33] is a statistical test that determines how the expected frequency versus the observed frequency, used to test the hypothesis, can be useful for investigating if one of the categorical variables depends on another. The Chi² test can also be employed as a feature selection method, primarily within the context of classification problems with categorical data. It is primarily used to sound the same as import factors with regard to how a factor is regarding to a locate variable. This results in a Chi-square statistic, which typically compares the observed (observed values) with what should be expected (expected values) in the data according to the statistics, using it to provide a pass/fail to whether this number is statistically significant or not. It is an indicator of how significant is the feature of interest when compared between the feature of interest and the target variable; with a high Chi-square value marking high dependence between between the feature of interest and the target variable, and low values mean that the feature and target variable are not dependent on the other, or that their relation with each other is minimal. It analyzes each property independently and determines which ones have a high impact on the target class.

The major benefit of this method is that it is quick and inexpensive to calculate; however, what it is really doing is looking for correlations between variables, checking to see how many times you observe the frequencies seen in the observed classes of other categorical classes. Therefore, it is very suitable for large-scale classification tasks, like text classification. Additionally, the simple procedure for the selection of important variables makes it easy to execute and interpret. A caveat, of course, is that the Chi-square test only applies to discrete events, so continuous variables need to be discretized into categorical bins before going through analysis (which generically is a limited approach). Moreover, where there are not enough samples in certain categories, the test may be less reliable, possibly resulting in inaccurate interpretations. Therefore scaling can still be a point of deficiency yet Chi-square is a fast and well-established technique commonly used for initial feature selection in high dimensionality classification problems. In Algorithm 2, pseudocode summarizing the steps of the Chi² used in this study is presented.

Algorithm 2 Chi^2 Method [33]

Input: Feature set $X = \{x_1, x_2, ..., x_n\}$, target variable Y, number of selected features k
Output: Selected feature subset S
Initialize $S \leftarrow \emptyset$, $R \leftarrow X$
for each $x_i \in R$ **do**
    Compute observed frequency O_{ij} for each category of x_i and class in Y
    Compute expected frequency E_{ij} assuming independence
    Compute Chi-Square score:

$$\chi^2(x_i, Y) = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

end for
Sort features by descending χ^2 score
Select the top k features into S
Return S

Feature selection strategies that use to extract important information from high dimensional data have a great impact on the classification performance for early diagnosis problems in other diseases. Specifically in identifying serious health concerns (such as cardiovascular diseases [34], malaria [35], and lung cancer [36]), the elimination of redundant or weakly correlated variables from the input dataset improved the model's accuracy and decreased the computational cost. Cardiovascular diseases were classified using methods based on ensemble classifiers, where it was used attribute pre-screening that reduced the class imbalance problem to obtain uniform prediction results. Similarly, previously in based deep learning models for malaria diagnosis, pre-processing routines were integrated into the architectures, i.e. CNN (Convolutional Neural Networks) and autoencoders, that eliminated background information, or white pixels, from analysis within the model enabling it to target biological relevant motifs. When diagnosing lung cancer, the support vector machines in the classification phase not only decreased the training time but also increased the precision of the model. And yet in specific applications they have changed how systems not only become easier but ultimately stronger, both in the model performance, as well as the clinical applicability.

3.2.1.3 Analysis of Variance (ANOVA)

ANOVA [37] is a type of parametric statistical test that is employed to identify mean differences between three or more samples. It serves primarily to evaluate relationships among variables by computing a ratio of between-group variance to within-group variance as potentially conditioned on a continuous variable. ANOVA works by measuring the difference in variances of each feature across classes and calculates how much that trait contribute towards classification; the higher the difference, the more importance it has, features with low differences could be excluded. The one-way ANOVA assists you in evaluating the impact of one independent variable, whereas you can utilize multi-way ANOVA (Two way or N-way) for verifying multiple independent variables with each other at the same time. The main statistic is the F-value, defined as the ratio of the between-group to within-group mean squares. The p-value provides a level of statistical significance, as determined against the F-value, which shows that the means by which groups are compared with each other contest significantly from one another. P-values (< 0.05 typically) suggest a statistically significant difference from at least one group mean, vs random differences as a function of noise.

The excellent aspect of ANOVA is that it allows comparing multiple groups as part of one single test, decreasing the failure rate from many pairwise comparisons. Some statistical assumptions are needed for ANOVA model results to be valid and reliable such as normality and homogeneity of variance. If these assumptions are not satisfied, the results may become unreliable, and alternative non-parameteric methods such as the Kruskal-Wallis test should be explored. Additionally, post-hoc tests (e.g., Tukey, Bonferroni) can be used to identify which group means are significantly different from each other. Over the years, ANOVA has been used extensively in experimental studies but it has also been successfully applied in areas including text classification, biomedical studies, and the social sciences as it helps to construct an efficient model through the identification and removal of similar or non-informative features. In Algorithm 3, pseudocode summarizing the steps of the ANOVA used in this study is presented.

Algorithm 3 ANOVA Method [37]

Input: Feature set $X = \{x_1, x_2, ..., x_n\}$, target variable Y, number of selected features k
Output: Selected feature subset S
Initialize $S \leftarrow \emptyset$, $R \leftarrow X$
for each $x_i \in R$ **do**
    Compute Between-Class Variance:

$$S_B = \sum_c N_c(\mu_c - \mu)^2$$

    Compute Within-Class Variance:

$$S_W = \sum_c \sum_{i \in c} (x_i - \mu_c)^2$$

    Compute ANOVA F-score:

$$F(x_i) = \frac{S_B}{S_W}$$

end for
Sort features by descending F score
Select the top k features into S
Return S

Relevant feature selection is still a significant step in high-dimensional datasets and is reflected in research carried out with the aim of enhancing segregation performance in the medical diagnostics process. Newer models for common diseases such as breast cancer [38], heart disease [39], and COVID-19 [40] have adapted methods of attribute selection based on differences between classes to not only improve inference accuracy by discarding redundant data, but improve system efficiency and interpretability as well. For instance in a breast cancer diagnosis study, the success of distinguishing benign vs. malignant tumors was great by using a tool of support vector machine after applying analysis variance based pre-selection process. For example, in a different study focusing on the prognostics of heart disease, the automatic feature selection process implemented with based variance technique not only exposed information on inter-class dissimilarities but also avoided model congestion with useless attribute. In a deep learning analysis of COVID-19 cases based on chest X-ray images, the selected set of features using variance-based feature elimination helped eliminate image-based noise and increased accuracy by concentrating the attention of the model to clinically

important patterns. ANOVA-based feature selection methods shows that instead of just statistical significance of ANOVA-based feature selection methods, they can also directly help to reliability of clinical decision support systems.

3.2.1.4 Kruskal-Wallis

The Kruskal–Wallis test [41] is a non-parametric statistical test for independent samples of different sizes or units to test the distribution or equality of population med. The Kruskal-Wallis test is used when the assumptions of normal distribution or homogeneity of variance are violated, so it's a nonparametric alternative to ANOVA. This test ranks the observations in the dataset and evaluates the ordinal distributions between groups also evaluating whether these ordered means are statistically means are not different. It works particularly well when dealing with ordinal data, where it is commonly used as a feature selection technique. Kruskal-Wallis helps to identify specific features that have major differences among classes according to their rank distributions which subsequently leads to better model performance through selecting informative features and discarding noise features.

The test is the same as the hypothesis (H_0) that under the null hypothesis H all groups have the same distribution and so do not differ from each other at all. The H statistic derives from the extent of the ordinal differences in groups. A high H with p-value smaller than a predefined value (e.g., $p < 0.05$) indicates that at least one group is statistically different. Nevertheless, the test does not reveal which groups differ from one another, and thus post-hoc analyses (e.g., Dunn's test) need to be conducted. One key advantage is its robustness: Kruskal-Wallis does not require normally distributed data, works well with small samples, and is resistant to outliers. However, it makes the assumption of similar distributional shapes across groups and does not directly assess magnitude of differences. When group distributions differ, this can make interpretation challenging. In summary, despite these limitations The Kruskal-Wallis is a popular and useful method for assessing group differences, particularly in the fields of medicine, biology, social sciences, and survey-based research when parametric assumptions cannot be met. In Algorithm 4, pseudocode summarizing the steps of the Kruskal-Wallis used in this study is presented.

Algorithm 4 Kruskal-Wallis Method [41]

Input: Feature set $X = \{x_1, x_2, ..., x_n\}$, target variable Y, number of selected features k
Output: Selected feature subset S
Initialize $S \leftarrow \emptyset$, $R \leftarrow X$
for each $x_i \in R$ **do**
    Rank all values of x_i across all classes in Y
    Compute mean rank for each class:

$$R_c = \frac{1}{N_c} \sum_{i \in c} R_i$$

    Compute Kruskal-Wallis statistic:

$$H = \left(\frac{12}{N(N+1)} \right) \sum_{c} N_c R_c^2 - 3(N+1)$$

end for
Sort features by descending H score
Select the top k features into S
Return S

In the context of the analysis of medical and biomedical data, they become particularly important, especially as non-parametric methods provide the ability to detect meaningful differences among groups when the assumptions of parametric methods are not valid or group distributions are unknown. The Kruskal-Wallis test is considered a robust and flexible attribute selection method as it is a non-parametric ranking based structure. In fact, this has been suggested in a study [42], where the inclusion of this test in feature ranking along with ANOVA contributed positively to machine learning classifier performance; it was reported that classification performance of decision trees, support vector machines, KNN greatly improved. Compared to more traditional approaches to genetic association analysis, especially regarding genetic variations associated with disease, the generalized Kruskal-Wallis test is able to provide considerable strength and meaning even in situations where there is uncertainty of data structure group membership [43], thus allowing easier interpretation to genetic variance. Alternatively, the Neutrosophic Kruskal-Wallis H test used [44] in the COVID-19 data analysis has allowed a more robust and inclusive assessment of the pandemic dataset through the generalization of the classical method in the presence of fuzzy, incomplete, or contradictory data. The introductory part

starts with the significance of the Kruskal-Wallis test, as one of the most used nonparametric hypothesis tests in applied fields ranging from genomics and bioinformatics through feature selection to group comparisons and analyses under uncertain conditions.

3.2.2 Metaheuristic Algorithms for Feature Selection

In this subsection, we explain metaheuristic feature selection methods that are grey wolf optimization, harris hawk optimization, particle swarm optimization, and whale optimization performed in this thesis.

3.2.2.1 Grey Wolf Optimization (GWO)

GWO [45] is one among the strongest nature-inspired metaheuristic algorithms and the results exhibit excellent performance on high dimensional data sets in feature selection. Inspired by the natural hierarchical hunting process in a gray wolf pack with alpha(the best solution), beta, and delta (the 2nd and 3rd best solutions) being the hierarchy and the other (omega) helping to explore the space, this technique works by initializing a gray wolf population on a search space. But because the algorithm is the result of the cooperation of all the pack members, such a mechanism allows to maintain an effective balance of exploration and exploitation: as a single pack all members search for the global optimum while avoiding the local minima traps.

GWO is most frequently utilized for reducing irrelevant or redundant attributes after finding the optimal subset of features to enhance model performance. It benefits from a flexible structure that can capture complex relationships between variables which traditional approaches may miss, making it particularly well-suited for high-dimensional and nonlinear problems. Also, GWO is widely used for engineering, artificial intelligence, and general optimization problems, including parameter tuning. A major advantage is its ability to lower computational cost at the same time as selecting relevant features. This will help in creating increasingly quick-performing, more durable, and more generalizable models as they focus on

Algorithm 5 The Grey Wolf Optimization (GWO) [45]

Input: Feature set $X = \{x_1, x_2, ..., x_n\}$, target variable Y, number of wolves N, maximum iterations max_Iter, hold-out partition HO
Output: Selected feature subset S
Initialize population X with random binary values $(0, 1)$
Evaluate fitness of each wolf using the fitness function
Sort wolves by fitness and select α, β, and δ wolves
Initialize iteration counter $t = 1$
while $t \leq max_Iter$ **do**
    Compute the adaptive coefficient $a = 2 - 2(t/max_Iter)$
    for each wolf i **do**
        for each feature dimension d **do**
            Generate random coefficients C_1, C_2, C_3
            Compute distances:

$$D_\alpha = |C_1 X_\alpha(d) - X_i(d)|$$

$$D_\beta = |C_2 X_\beta(d) - X_i(d)|$$

$$D_\delta = |C_3 X_\delta(d) - X_i(d)|$$

            Compute updated positions using transfer function:

$$Bstep = \frac{1}{1 + \exp(-10(D - 0.5))}$$

            Update wolf position based on probability
        end for
    end for
    Evaluate fitness of each wolf
    Update positions of α, β, and δ wolves
    Store best fitness value in convergence curve
    Increment iteration counter $t = t + 1$
end while
Identify best feature subset S from α wolf
Return S

diluting the impact of multi-collinearity and reducing the opportunity of overfitment in the process. The performance of GWO, however, depends on the parameters' values and also on the initial population, so the correct choice and hyperparameter tuning must be done. Thus, GWO is not only one of the outstanding techniques for feature selection(s) as data preprocessing, but also an efficient approach for increasing the overall efficiency of complex applications. In Algorithm 5, pseudocode summarizing the steps of the GWO used in this study is presented.

The finding from some of the recent studies regarding the diagnosis of life threatening diseases including Heart diseases [46] [47] and types of cancer [48] prove that approaches based on GWO are effective in selecting features that effectively have deep impact on the classification algorithms performance. The settings of GWO were indeed used to bolster precision of exactness by removing the redundant features and structuring a narrower model avoiding the model from terrific overfitting in combination with support vector machines according to the output of coronary artery disease classification. Moreover, in another GWO-based system for heart disease early detection, it expedited convergence, and improved performance by means of population-based search strategies. Deep learning uses GWO and two-dimensional texture features in hybrid models that can reduce complex data with high performance, using color histopathology images of the patients with lung and colon cancer. In conjunction with the Mobile-EfficientNet model family, it shows the power to highlight clinically meaningful textural characteristics. The descriptions and results of these studies suggest that GWO is an effective feature selection method that both enhances accuracy and makes it easier to build models that can be incorporated into a clinical system and work with fewer and more efficient feature sets.

3.2.2.2 Harris Hawk Optimization (HHO)

One of the competent metaheuristic algorithms that have been produced for selecting the most significant attributes from a high-dimensional dataset is HHO. Although there is no specific behavioral mechanism by hawks, inspired by the hunting strategies of Harris hawks in nature, HHO searches the best solution in the solution space by simulating the collaborative behavior of hawks in a group.

HHO [49] is one of the effective metaheuristic algorithms that has been developed for the selection of the most important attributes from a high-dimensional dataset. Inspired by hunting habits and group behavior of Harris hawks in nature, HHO does not have a specific behavioral mechanism by hawks, searches for the best solution in the solution space simulating the social behavior of hawks in a group. Hawk searching algorithm starts with attribute subsets randomly generated, and when it uses different hawk

attack strategies against a target, which is the best attribute subset, and is improved iteratively. All this together leads to sufficient exploration of the search space and ultimately selection of relevant attributes. A significant aspect of this algorithm is its capability to balance exploration and exploitation phases of this algorithm help the algorithm avoid locally optimal solutions and led the algorithm to finally reach the global optimum. HHO has been successfully implemented to deal with engineering, artificial intelligence, and parameter tuning problems, particularly for a complex and high-dimensional optimization problem.

Since it is necessary to maintain a representative feature set, one of the major advantages of HHO is its reduced computational cost and speedup, so it is efficient in big data. This helps in reducing multicollinearity and also helps create better models since by removing redundant features it helps in building more stable/less generalizable models. HHO's search mechanism enables it to evaluate many alternative feature combinations in a limited time frame, which combats overfitting. Due to the stochastic nature of this algorithm, we do have varying results between different runs, thus fine-tuning of hyper-parameters is required. Overall, it can be seen that HHO constitutes a strong and powerful method for feature selection that can be applied at the data preprocessing stage resulting in the development of better-performing models with the least possible effect of redundant or noisy dataset features. In Algorithm 6, pseudocode summarizing the steps of the HHO used in this study is presented. In the medical dataset which is often complex and high-dimensional, the last few years have seen the emergence of a strong need for determining the appropriate feature sets of its dataset for people who wish to generate a classification model. In this context, the Harris Hawk Optimization (HHO) algorithm provides a powerful answer, particularly multidimensional, as it solves more than one goal simultaneously (accuracy, processing time, and model simplicity). Such HHO-based hybrid approaches integrated data-wise with deep neural networks in the diagnosis of malignant lymphoma cancer [50] facilitated the model's learning to get quicker and result in a less error rate by choosing the most significant and meaningful of the data sets comprised of a multitude of biomarkers. There is another study [51] in which an explosive search mechanism is added to the classical version of

Algorithm 6 The Harris Hawk Optimization (HHO) [49]

Input: Feature set $X = \{x_1, x_2, ..., x_n\}$, target variable Y, number of hawks N, maximum iterations max_Iter, hold-out partition HO
Output: Selected feature subset S
Initialize population X with random binary values $(0, 1)$
Evaluate fitness of each hawk using the fitness function
Identify the best hawk X_{rb} (rabbit position)
Compute the mean position X_μ of the population
Initialize iteration counter $t = 1$
while $t \leq max_Iter$ **do**
    for each hawk i **do**
        Evaluate fitness of hawk X_i
        if fitness of X_i is better than X_{rb} **then**
            Update X_{rb}
        end if
    end for
    Compute escaping energy $E_0 = -1 + 2 * \text{rand}()$ and $E = 2 * E_0 * (1 - t/max_Iter)$
    for each hawk i **do**
        if $|E| \geq 1$ **then** ▷ Exploration phase
            Choose a random hawk X_k
            Update position using stochastic rules
        else ▷ Exploitation phase
            Generate random J and r
            if $r \geq 0.5$ and $|E| \geq 0.5$ **then**
                Update position using soft besiege strategy
            else if $r \geq 0.5$ and $|E| < 0.5$ **then**
                Update position using hard besiege strategy
            else if $r < 0.5$ and $|E| \geq 0.5$ **then**
                Perform Levy flight-based update
            else if $r < 0.5$ and $|E| < 0.5$ **then**
                Perform combined Levy flight-based update
            end if
        end if
    end for
    Store best fitness value in convergence curve
    Increment iteration counter $t = t + 1$
end while
Identify best feature subset S from X_{rb}
Return S

the HHO algorithm to improve its classification success and parameter search capability for classifying COVID-19 patients based on severity of the infection. Finally, because the parallel multi-objective HHO algorithm [52] is used for the prediction of mortality associated with COVID-19, we were motivated to develop an

effective, personalized prediction mechanism for any prediction system (which is predominantly used for Intensive care decision support systems). For instance, covid prediction can be modeled as a multi-criteria optimization problem, i.e. instead of maximizing only one object, multiple performance criteria (accuracy, sensitivity, and computational burden) must be maximized at the same time. The available research suggests that HHO can be considered as a heuristic search algorithm and as a versatile and robust feature selection approach to be utilized in complex decision-making processes in clinical risk stratification.

3.2.2.3 Particle Swarm Optimization (PSO)

PSO [53] is also one of the well-established and well-performed metaheuristic algorithms, receiving a lot of attention in the field of feature selection problems, especially for the high-dimensional datasets. Based on the flocking of birds or fish schooling, PSO is, in essence, a population-based optimization tactic where all candidate solutions, referred to as particles, traverse the search space in the hopes of finding an optimal solution. A mix of both the particle's own best-known position and the best-known position of the swarm updates the particles. Updating particles in this manner reduces the chance that the swarm will become trapped in local optima and increases the chances that the swarm will find globally optimal solutions. PSO has proven to be broadening the hands of combination in both feature selection and parameters fine-tuning tasks, as well as having very successful performance in the fields of engineering and artificial intelligence.

From an academic perspective, PSO, especially its utility in optimising regression models, has several significant advantages one of which being that it lowers computation cost and hence increases the model performance by discovering a few features distributed in high dimensional spaces. It mitigates the risk of overfitting and addresses multicollinearity, making the models more robust and generalizable. Its randomised but adaptive searching process also helps in preventing overfitting. But, due to stochasticity in how the algorithm operates, particularly in how it initializes particle positions and velocities, it can be sensitive to hyperparameter configuration

and can produce different results in different runs. This work incorporates a Binary PSO-based variant. PSO is recognized as one of the state-of-the-art feature selection method and preprocessing techniques in recent years in data analysis workflows, as it supports reliability and robustness in processes. In Algorithm 7, pseudocode summarizing the steps of the PSO used in this study is presented.

Attribute redundancy can adversely affect model performance in high-dimensional medical data, especially at such complex high levels of structures as microarrays. As a result, defining suitable and informative features on clinical classification procedures needs to be recognized in terms of not only the precision but also the computing time. PSO is one of the most common methods in biomarker selection, and image-based diagnosis systems in the literature. For instance, in a PSO-based study focusing on microarray-based cancer diagnosis [54], the algorithm not only adapted to the binary structure enabled the removal of redundant genes from the classification model, thus achieving improved classification accuracy using only relevant biomarkers. In a deep learning based study [55] entitled as effective detection of the pneumonia in chest X-ray, an adaptive and collaborative approach for PSO is designed which filters a huge number of attributes obtained from the deeper layers by keeping focus on only the ones that contribute towards diagnosing and therefore enhancing the generalization capability of the model. Another study [56] on prostate cancer diagnosis based on PSO showed that reducing the amount of the microarray data from 126 to 18 by dimensionality reduction helped improve classification performance with significantly reduced training time. These examples show that, due to its universality and performance acceleration capabilities, the PSO algorithm can be applied not only with classic Machine Learning models, but also with deep learning systems and can help with the construction of the trustful, fast, and highly accurate diagnostic methods of diseases based on provided clinical data.

3.2.2.4 Whale Optimization (WO)

WOA [57] is an evolutionary algorithm inspired by nature which is used widely metaheuristic calculus that has been presented to investigate most relevant features by

Algorithm 7 The Particle Swarm Optimization (PSO) [53]

Input: Feature set $X = \{x_1, x_2, ..., x_n\}$, target variable Y, number of particles N, maximum iterations max_Iter, cognitive factor c_1, social factor c_2, hold-out partition HO
Output: Selected feature subset S
Initialize population X with random binary values $(0, 1)$
Initialize velocity V with zeros
Evaluate fitness of each particle using the fitness function
Set personal best position X_{pb} and global best position X_{gb}
Initialize iteration counter $t = 1$
while $t \leq max_Iter$ **do**
    Update inertia weight $w = W_{max} - (W_{max} - W_{min}) \times (t/max_Iter)$
    for each particle i **do**
        for each feature dimension d **do**
            Generate random numbers r_1, r_2
            Update velocity:

$$V_{i,d} = w \cdot V_{i,d} + c_1 \cdot r_1 \cdot (X_{pb,i,d} - X_{i,d}) + c_2 \cdot r_2 \cdot (X_{gb,d} - X_{i,d})$$

            Apply velocity limits V_{max}
            Compute transfer function:

$$TF = \frac{1}{1 + \exp(-V_{i,d})}$$

            Update position based on probability
        end for
        Evaluate fitness of particle X_i
        if fitness of X_i is better than $X_{pb,i}$ **then**
            Update $X_{pb,i}$
        end if
        if fitness of $X_{pb,i}$ is better than X_{gb} **then**
            Update X_{gb}
        end if
    end for
    Store best fitness value in convergence curve
    Increment iteration counter $t = t + 1$
end while
Identify best feature subset S from X_{gb}
Return S

examining the high-dimensional dataset and eliminate highly correlated features. It is specifically done to preserve features that provide maximum discord information with respect to the target variable while removing features that contribute to redundant overlap. The proposed WOA is inspired by the spiral hunting method and bubble-net

feeding method of the humpback whales that simulates the process of whales encircling and squeezing their prey. We will also demonstrate how to leverage the dynamics of GWO to do feature selection where in the context of GWO, each whale in the population becomes a candidate solution and updates its current position in feature space iteratively according to the best-known solution and the direction of a random search in feature dimensions. This mechanism enables a balance between exploration and exploitation phases for the algorithm to lead itself to the global optimum as long as local minima are avoided.

The significant optimization power of WOA and its strong simplicity in dealing with any kind of data structure leads to its broad-scale application in engineering, artificial intelligence, and parameter tuning. Feature selection is a process that helps reduce the number of features in high-dimensional data thus decreasing computational cost and increasing model accuracy. Besides, it also reduces multicollinearity, which results in more stable and generalizable models and its flexible structure decreases the danger of overfitting. However, the performance of WOA is greatly influenced by parameters settings, and improper parameters settings would trap in searching strategies into local optimum. In general, WOA provides a powerful and strong feature selection strategy that substantially enhances the data preprocessing and facilitates the building of effective predictive models. In Algorithm 8, pseudocode summarizing the steps of the WOA used in this study is presented.

The problem of selecting the most informative features to improve model performance through efficient feature selection is prominent in the field of medical data analytics, especially in the case of high dimensional and noisy data. In this respect, the Whale Optimization Algorithm (WOA) has shown up as one of the most robust metaheuristics in medical classification problems, due to its superior search capability and flexible setting potential. In an empirical study [58] on heart disease prediction, has compared the performance of WOA for different configurations of parameters while confirming that the algorithm enhances classification accuracy and optimizes processing time rendering the model appropriate for practical applications. Another study [59] conducted involved finding out the low contribution features to the

Algorithm 8 The Whale Optimization (WO) [57]

Input: Feature set $X = \{x_1, x_2, ..., x_n\}$, target variable Y, number of whales N, maximum iterations max_Iter, hold-out partition HO, constant b
Output: Selected feature subset S
Initialize population X with random values in range $[0, 1]$
Evaluate fitness of each whale using the fitness function
Identify the best whale X_{gb}
Initialize iteration counter $t = 1$
while $t \leq max_Iter$ **do**
    Compute the adaptive coefficient $a = 2 - t \times (2/max_Iter)$
    for each whale i **do**
        Generate random numbers $A = 2 \times a \times \text{rand}() - a$, $C = 2 \times \text{rand}()$, $p = \text{rand}()$, $l = -1 + 2 \times \text{rand}()$
        if $p < 0.5$ **then**
            if $|A| < 1$ **then** ▷ Encircling prey
                Update position using:

$$X_i = X_{gb} - A \times |C \times X_{gb} - X_i|$$

            else ▷ Exploration phase
                Select a random whale X_k
                Update position using:

$$X_i = X_k - A \times |C \times X_k - X_i|$$

            end if
        else ▷ Bubble-net attacking phase
            Update position using:

$$X_i = |X_{gb} - X_i| \times \exp(b \times l) \times \cos(2\pi l) + X_{gb}$$

        end if
        Apply boundary constraints on X_i
    end for
    Evaluate fitness of each whale
    Update the best whale X_{gb}
    Store best fitness value in convergence curve
    Increment iteration counter $t = t + 1$
end while
Identify best feature subset S from X_{gb}
Return S

diagnosis of COVID-19 cases, that is, features were identified to analyze which features are not contributing much towards the diagnosis the COVID-19 cases using the WOA feature selection in order to eliminate them and build a lesser number of considerable, highly interpretable yet accurate models. In addition, the WOA's

adaptations for conducting the multi-class skin cancer diagnosis problem [60] made the full models insensitive to the class balance for the multi-class type and able to produce results that can be integrated into clinical decision processes, and improved the separate area under the receiver operating characteristic curve for each type of multi-class cancer. These applications reflect that WOA is not only one of the optimization algorithms that draw attention with its optimization performance, but also can be a commonly used optimization algorithm to form stable and efficient decision support systems in health data.

3.3 Machine Learning Models

We perform several types of machine learning classifiers to predict patients with gallstones, and in this section, we briefly summarize these classifiers that are Logistic Regression, Naive Bayes, Decision Trees, K-Nearest Neighbors, Support Vector Machines, Gradient Boosting, AdaBoost, Random Forest, and Multilayer Perceptron.

Logistic Regression is indeed very simple to implement and provides good results on linearly separable data, but the above procedure does not, however, work well with complex nonlinear associations unless there are transformed features available. Because this methodology is straightforward, Naive Bayes is simple enough to be computationally efficient and fast, but the need for features independence hurts its performance on real-world data. A decision tree can be visualized and is easy to interpret, but the tradeoff is less generalizability which can cause overfitting. K-Nearest Neighbor (KNN) is an interpretable algorithm and does not have a training phase but has the disadvantages of having a high prediction time, sensitivity to irrelevant features, and sensitivity to data scaling.

Support Vector Machines (SVM) are very strong classifiers that can create complex decision boundaries, but they can be computationally intensive for large datasets. Implementations such as Gradient Boosting and AdaBoost both enhance the accuracy of predictions by amalgamating weak learners, however may have a slower training time

and be sensitive to noise in the data. Random Forest, are both less prone to overfitting because of average over multiple trees, and more valid, but at the cost of a bit of interpretability. The function of the MLP (Multi-layer Perceptron) has been formed neural network that is well-suited for complex nonlinear relationships that describe high-dimensional data, although they may require significant computational resources and tuning of parameters in order to prevent overfitting and achieve convergence.

We also took a deep dive into how the performance of these algorithms varies with feature selection. In particular, find the highest performing models using the least amount of features. It was also seen that the simple models like the Logistic Regression and the Naive Bayes were still performing decently with very few features, while the complex models like SVM, Gradient Boosting, and MLP needed more features to be able to perform to their full capacity. Decision Trees and Random Forests performed well despite generating only a minority of features providing support for its implementation for feature selection. This explains and gives clear feedback on how to choose the appropriate model for a given application and what should or should not work there based on the number of dimensions in the corresponding feature space and the data structure.

In this study, the hyperparameter optimization of machine learning algorithms is performed using the default parameters of Python. It does not mean that such an idea is wrong, and to prefer it is due to the fact that this overview allows one to focus on comparing for the standard performance of the algorithms instead of the very comparison for greedy searches of hyperparameter values. This enabled a better model selection and was a reason that differences in performance for algorithms were easily observed. We can further check whether hyperparameter tuning towards the performance of the model (grid search, bayesian optimization) improves accuracy.

3.3.1 Machine Learning Algorithms

We executed various classifiers for predicting gallstones patients and briefly summarize these classifiers in this section. All the machine learning classifiers

employed in this study were summarized in the pseudocode of the algorithms as in 9.

Algorithm 9 Pseudo Code of Machine Learning Classifiers

Input: Dataset D, target variable Y, test set ratio r, random state s
Output: Evaluation results for each classifier
Load dataset D from source file
Remove unnecessary rows and columns from D
Extract target variable Y from D
Remove Y column from D
Convert D and Y to NumPy arrays: $X \leftarrow D$, $y \leftarrow Y$
Convert values in X and y to float type
Split data into training and test sets:

$$X_{train}, X_{test}, y_{train}, y_{test} \leftarrow \text{train_test_split}(X, y, \text{test_size} = r, \text{random_state} = s)$$

Initialize a list of classifiers with their default parameters:

$$models = \{(name_1, model_1), (name_2, model_2), ..., (name_n, model_n)\}$$

for each $(name, model)$ in $models$ **do**
    Train the model using X_{train}, y_{train}
    Predict labels: $\hat{y} \leftarrow model.predict(X_{test})$
    Compute and print evaluation metrics:

- Accuracy: accuracy_score($y_{test}, \hat{y}$)
- AUC: roc_auc_score($y_{test}, \hat{y}$)
- Classification Report: classification_report($y_{test}, \hat{y}$)
- Confusion Matrix: confusion_matrix($y_{test}, \hat{y}$)

end for

3.3.1.1 Logistic Regression Algorithm

LR (Logistic Regression) [61] is among the most used supervised machine-learning algorithm for binary classification problems. While other algorithms model an outcome based on a direct linear relationship between the input features, and the output classes, LR applies the sigmoid function to the linear combination of features so that it is represented as a probability value between 0 and 1. This is given is sigmoid**3.1**, where z is a weighted sum of the input features. LR can then classify data according to a fixed threshold (often 0.5), creating a linear decision boundary in the feature space. In fields like medicine and finance, where the interpretability of the model is of extremely high importance, one of the advantages of LR being an interpretable

model really shines through: The coefficients of the model can be easily understood as log-odds of the datapoint output. Moreover, it is computationally inexpensive, simple to implement, and acts as a strong baseline model in classification problems. LR is based on a relationship being linear around the log-odds of the target variable in relation to independent predictors, which is not well suited for more complex or nonlinear distributions of the data.

$$g(z) = \frac{1}{1 + e^{-z}} \tag{3.1}$$

Here we used Logistic Regression as a baseline classifier to study performance on the preferred dataset. Input features were also normalized before training to ensure the efficient optimization of the trained model. The model was trained using the default logistic loss and no regularization technique. Standard classification metrics were used to evaluate the performance of the classifier, providing a general insight into the classifier's performance, such as accuracy, precision, recall, and F1-score. Although the Logistic Regression does not necessarily perform very well with cases in which the target variable has nonlinear distribution or significant outliers, the simplicity and interpretability of the Logistic Regression still make it a solid baseline by which the performance of the more complex models to be explored in the next sections should be compared against.

Logistic regression is a potent classification technique characterized by a relatively simple structure that was employed in models for the development of medical decision support systems to balance accuracy and interpretability. In clinical situations the motivation to learn cause and effect relationships of diseases is somewhat at the leading edge to this, which is why the interpretable coefficients of logistic regression help explain the output of the model for physicians and add to the impression of reliability. Logistic regression analysis revealed that disease-related variables

significantly affected the success of the classification of thyroid cancer diagnosis in their study [62]. In particular, the statistically significant biomarker areas for early diagnosis were calculated according to this type of analysis. Likewise, studies that performed classification of brain tumors using MR spectroscopy data found that their method [63] improved accuracy by distilling complex signal data to a format that the model could process, while providing an explainable model on which clinical decisions could be based. In a separate analysis [64] of cancer risk factors, logistic regression provided not only a classification tool, but specific determinants that were inversely correlated to disease, thus aiding in the implementation of preventive public health. Representing these points, logistic regression is an elegant, powerful, model-simple, and clinically translatable approach to diagnostic and risk modeling processes.

3.3.1.2 Naive Bayes Algorithm

Naive Bayes (NB) [65] is a popular supervised machine learning algorithm that is simple but effective for classification tasks. It is based on Bayes' Theorem and applies the naive assumption that all features (inputs) are conditionally independent given the class label. This strong assumption greatly simplifies the computations and gives the model the power to compute probability estimates very efficiently, even on high-dimensional datasets.

NB is a very well-known supervised machine learning algorithm that can be simple yet effective when it comes to the classification task. Simply put, it is based on Bayes' Theorem and the strong assumption of independence between features (inputs) given the class label. This stiff assumption simplifies calculations a lot while lending the model the ability to learn probability estimates quite effectively, even for high-dimensional datasets. Even though the assumption of independence is unrealistic in numerous real-world scenarios, Naive Bayes often achieves competitive and robust performance, especially in tasks like text classification, medical diagnosis, and spam filtering. For a given set of features, it computes the posterior probability of each class and classifies the observation to the class that has the highest posterior probability. Before we rush

to Naive Bayes classifiers we need to see the mathematical formulation of Bayes' Theorem, and in its most general form it can be written as:

$$P(C_k|X) = \frac{P(X|C_k)P(C_k)}{P(X)} \tag{3.2}$$

where:

- $P(C_k|X)$ is the posterior probability of class C_k given the features X,
- $P(X|C_k)$ is the likelihood of observing X given class C_k,
- $P(C_k)$ is the prior probability of class C_k,
- $P(X)$ is the probability of X.

In more detail, the Gaussian Naive Bayes variant was used, because features in the explained dataset are critical and are assumed to be normally distributed within each class. The features were scaled and preprocessed before training, to be compatible with the probabilistic structure of the model. They used the algorithm for baseline classifiers, and their performance was evaluated with a set of metrics (accuracy, precision, recall, f1-score). Due to the assumptions made in Naive Bayes where it considers all features independent from each other, thus ignoring the interaction between features, Naive Bayes is a very fast classifier with lower computational complexity and gives reasonably good performance on many real world classification tasks. It also serves as a strong baseline against which more complex machine learning models covered in the ensuing sections can be compared.

In medical classification problems, fast, low-cost, and accurate models are of great importance, particularly in situations requiring early diagnosis. The structure based on statistics and its easy-going assumptions have made the naive bayes algorithm as one of the popular techniques for such based clinical applications. The research [66] focused on breast cancer detection showed that the algorithm reached very high levels

of accuracy even with having limited and largely categorical information, which shows strong applications to the field of early diagnosis. In many recent models [67] applied to the diagnosis of Alzherimer's disease, a high accuracy was achieved for predictions based solely based on clinical findings by the algorithm's great generalization ability, showing that Naive Bayes is a competitive method in the classification of neurodegenerative diseases. In another study [68], the algorithm trained in demographic and radiographic data on cases with bone tumors, for instance, successfully utilized numerical and categorical variables in differential diagnosis. The published papers demonstrate that Naive Bayes is commonly utilized in the clinical area with not just high accuracy, but also because it is interpretable, computationally inexpensive, and proper for small data sets.

3.3.1.3 Support Vector Machine Algorithm

SVM (Support Vector Machine) [69] is a powerful supervised machine learning algorithm used primarily for classification problems, particularly for high dimensional, and complex data. Its fundamental aim is to find a hyperplane in N-dimensional space that distinctly classifies the data points of different classes as far as possible by maximizing the distance between them (the margin between classes) to the nearest points from each class, called the support vectors. The idea is that when the data to be classified are not linearly separable in the original space, a kernel function can be used to implicitly map the data in a higher-dimensional feature space where the data may be linear separable. SVM is thus very flexible, as it can produce both linear and nonlinear decision boundaries based on the used kernel. The optimization problem in SVM is given as:

$$\min_{\mathbf{w},b} \frac{1}{2}|\mathbf{w}|^2 \quad \text{subject to} \quad y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1, \quad \forall i \tag{3.3}$$

where:

- w the weight vector perpendicular to the decision boundary,
- b is the bias term,
- $y_i \in \{-1, +1\}$ are the true class labels,
- x_i are the input feature vectors.

In the study, the SVM algorithm is one of the classification models to evaluate the performance of the selected datasets. Feature scaling is critical to algorithms such as SVM, therefore prior to model training all features were standardized. We applied the model with default parameterizations employed on the dataset without the selection of any kernel or tuning of hyper-parameters. The classification performance was evaluated using standard metrics such as accuracy, precision, recall, and F1 score. Only training with SVM, it has a good performance, especially in high-dimensional space, but the effectiveness is limited by the choice of kernel and parameter settings. However, because of its ability to produce clear decision boundaries, it offers a useful benchmark model against which to compare other classification methods we will describe later in this chapter.

For the task of classifying real, challenging, high-dimensional, noisy data sets with complex structures, such data sets are becoming more prevalent in medicine, prompting the need for accurate and generalizable classification models. Even so, the SVM (Support Vector Machine) algorithm is commonly favored in clinical classification problems due to its potential to address high-dimensional data, its potential which clearly create decision boundaries, and its reliable performance regardless of the small sizes of data sets. For instance, in a research [70] for heart diseases prediction, the SVM algorithm produced very high accuracy for healthy and sick individuals classification, and also demonstrated strong performance based on sensitivity and specificity values. In another study [71] for early diagnosis of Parkinson's disease, SVM achieved good separation of classes at the symptomatic stage, boosting the chance for early intervention. A breast cancer diagnosis [72] using an SVM-based ensemble model achieved higher accuracy by concentrating on a smaller quantity of predictive features; it also solidified the clinical applicability of the

model. These examples demonstrate that SVM has versatile advantages of high accuracy, precision of class discrimination and success of generalization, especially contributing to biomedical classification problems.

3.3.1.4 Decision Tree Algorithm

DT (Decision Tree) [73] is one of the most well-known supervised machine learning algorithms that is used for classification or regression. This works by splitting iterator data into smaller and smaller sets, based on the values of different features, and creating a tree of decisions. At each step, the algorithm chooses those features with the highest information gain (or other splitting criteria like Gini impurity) in order to keep child nodes as pure as possible. This greedy, top-down approach is called recursive binary splitting. One of the main benefits of decision trees is the ability to explain. They are not only easy to implement, but the tree structure offers a simple and intuitive visualization of the decision-making process. This makes them interpretable even for those lacking an advanced education in mathematics, as the logic behind the model often mirrors human intuition. Additionally, very little data preprocessing is required with Decision Trees, and they can handle both numerical and categorical data seamlessly. The generalization can be mathematically defined in terms of the information gain (IG) given by:

$$IG(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{3.4}$$

where:

- S is the dataset,
- A is a feature,
- S_v is the subset of S where feature A has value v,

- $Entropy(S)$ is the entropy of the dataset S.

A Decision Tree classifier was also used as one of the baseline models for benchmarking on this dataset in this study. As Decision Trees is insensitive to feature scaling we did not perform normalization on the data. Default Settings: trained with default settings, no pruning, depth limit. Standard classification metrics such as accuracy, precision, recall, and F1-score, were used to evaluate the performance of the model. Decision Tree, although capturing the true meaning of simplicity and explainability, is the most overfitted algorithm due to their deep depth causing the trees to memorize the training data. This limitation was addressed in this analysis through the interpretation of results, but in future work, regularization techniques like pruning or max-depth control may also be applied. Nonetheless, Decision Trees provide a quick and transparent classification methodology and are a good baseline to compare with the more sophisticated models that will be described in later sections.

It is essential to create models that are both accurate and easily interpretable in medical decision support systems in order for the physicians to directly gain from the model outputs. Is there a practical aspect? At some stage in their journey, Decision Tree algorithms are overwhelmingly preferred in clinical applications due to their visual explainability and intuitive structure. In heart disease prediction, when statistical classification models applied to the clinical data, decision trees proved a strong helper due to their easily understandable classification structure using the most relevant features providing important insight into major factors implicated in heart diseases [74] that assist the physician to determine an outcome. The different decision tree techniques implemented in the prediction of liver diseases [75] improved the sensitivity of the models and proved particularly successful in discriminating the risk levels of the disease. Such systems based on artificially-intelligent developed explicitly for providing correct and early diagnosis for the first stage indications of Parkinson's disease [76] comprising diagnostic modeling correctly enough for the upper level decision making purpose to make sure about the start of such ailment yet being quite simple enough from the dangers perspective based on experiences likewise of a typical ascribes of human. All these studies highlight that Decision Tree

algorithms provide a very strong alternative for medical classification problems in terms of not just performance, but also in terms of transparence, and clinical interpretability.

3.3.1.5 K-Nearest Neighbor Algorithm

KNN (K-Nearest Neighbors) [77] is a traditional machine learning algorithm used for classification tasks. An instance-based method like K-Nearest Neighbors (KNN), on the other hand, is an example of a non-parametric approach, because it does not impose a parametric functional form on the input features and response, but there are many further and less advanced yet very useful distinctions, where KNN is an example of an instance-based method, a probabilistic or model-based method (KNN algorithm is very interesting because it doesn't learn from the training data at all. It does not learn a model, but simply marks new instances based on how similar they are to those already in the teaching set. We find this unique characteristic of KNN worth exploring further.). KNN is an instance-based learning algorithm, which classifies a new target based on the majority class of its k closest neighbors based on a distance measure like Euclidean distance. Since KNN has no learning phase (no training phase by conventional means), all computational effort is postponed until its time for prediction; KNN is therefore a lazy learning algorithm. The Euclidean distance that represents a mathematical formulation for similarity measurement between two data points $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_n)$ is given by:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3.5}$$

where:

- $d(x,y)$ is the distance between points x and y,
- x_i and y_i are the values of the ith feature of each point,

- n is the total number of features.

The KNN algorithm was used in the study as a base classifier to assess performance on the dataset chosen. As KNN being a distance-based algorithm, feature scaling has been helpful to bring all the attributes on similar scales before to be added to the computation of distance to be trained. More specifically, k was fixed and the Euclidean distance metric was employed for the neighbor selection. A standard evaluation of classifier performance metrics: accuracy, precision, recall, and F1-Score were also examined. KNN is simple, intuitive, and straight forward to implement, however has some drawbacks, including high computational cost at prediction time, as well as being sensitive to irrelevant and highly correlated features. Nonetheless, KNN's simplicity and reasonable performance make it a good basis to benchmark performance against more complicated classifiers presented later in the tutorial.

In based classification healthcare systems, the non-parametric methods highly rely on the distribution and hence proves to be beneficial. In that context, the KNN algorithm has been found to be widely used in medical diagnosis systems in order to benefit from well known pattern recognition ability and its simplicity. The use of KNN algorithm based similarity-based classification approach in their study [78] to predict heart diseases, and reported an effective search approach differentiating the sick and healthy individuals and consistent results among KNN algorithms with a much greater focus towards maintain class balance. A fuzzy KNN-based model built for the detection of Parkinson's disease [79] provided more accurate results in the assessment of uncertain symptoms and thus helped to advance the early diagnosis. Classifying diabetes based on clinical features through measurement of distance between patients [80] gave results compliant to clinical findings in a separate clinical study for diabetes diagnosis, and the relatively simple model makes it more suitable for near-time implementations. KNN implementation with classical and hybrid solutions proves that the KNN algorithm is adaptive to diverse medical classification structures and also potential to enhance clinical decision support system's significantly incorporating interpretable and easy-to-deploy solutions.

3.3.1.6 Gradient Boosting Algorithm

GB (Gradient Boosting) [81] is one of the strongest and widely used ensemble learning techniques that can be applied for supervised machine learning problems, primarily regression and classification. Basics of Gradient Boosting In essence, Gradient Boosting is the method of solving predictive problems by finding the combined prediction from multiple base learners, in most cases decision trees. Gradient Boosting builds these weak learners sequentially, where each new learner attempts to correct the errors of the previous learners. The model gets better with every iteration after it has performed heuristic search over the loss function. The model is updated at each stage m as follows:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{3.6}$$

where:

- $F_m(x)$ is the updated ensemble model after m iterations,
- $F_{m-1}(x)$ is the ensemble model from the previous stage,
- $h_m(x)$ is the new weak learner trained on the residuals (errors),
- γ_m is the learning rate or step size controlling the contribution of $h_m(x)$.

This study utilized the Gradient Boosting Algorithm as one of the primary classifiers for assessing the model performance on the designed dataset. The features were preprocessed, and scaled where applicable, before applying the models. Base learner was set to default decision tree, with no manual tuning of hyperparameters. We used standard classification performance metrics like accuracy, precision, recall, and F1-score for model evaluation. Gradient Boosting is widely used for structured data due to its high predictive power and flexibility with respect to nonlinearities. However,

it is also susceptible to overfitting, if we don't control either the number of trees or the depth of all trees. Nevertheless, Gradient Boosting proved a still competitive algorithm and one that is often more effective than simpler models, and for that reason, it served as an interesting baseline in this case, as its performance would contribute to perceiving the efficacy of the suggested feature selection and modeling approach.

In medical classification, especially in cases with high dimensional data, there is an even greater need of powerful ensemble methods to learn these complex patterns and find high accuracy and very small but significant difference. Currently, the Gradient Boosting strengthens its performance by boosting weak classifiers consecutively and its appropriate predictive capabilities will yield positive outcomes in clinical diagnosis systems. Existing study [82] conducted for heart diseases detection with the Data-based technique of Gradient Boosting show higher accuracy and dimensionality information, and the early detection benefit of the risky individuals. A different study [83] aimed at diagnosing the audio form of Parkinson's Disease classifiers was able to effectively differentiate subtle vocal differences and show discrimination of multiple measures of the presence of the disease. Similarly, in this genetic (SNPs) analysis one of the powerful approaches was Gradient Boosting Tree as addiction study [84] biomarker discovery it aims at prediction of diseases from the given input and works quite efficiently for developing personalized health solutions So through these studies it was demonstrated that Gradient Boosting is a robust approach not only due to the high success with the classification, but because of the ability of adaptation to different types of data, learning complex patterns as well as their integrability in clinical applications.

3.3.1.7 Adaboost Algorithm

AB (AdaBoost) [85] is an ensemble learning algorithm of the nomothetic type which enhances the accuracy of weak learners by assembling a finite number of classifiers into a single strong classifier. It does so by sequentially training a collection of weak learners, which are typically decision stumps (one-node decision trees), where each subsequent classifier focuses more on those instances that were misclassified by the

previous classifiers. The basic principle is that AdaBoost assigns a weight to each of the training samples, and updates them and their weights every iteration by increasing the weights of wrongly classified samples and reducing the weights of correctly classified ones. So later learners focus on the more tricky cases. The final output is taken as the weighted average of all weak classifier predictions. The objective loss function to be minimized by the AdaBoost algorithm is the exponential loss function, and the prediction rule associated with this type of loss function can be formulated as:

$$H(x) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m h_m(x)\right) \tag{3.7}$$

where:

- $H(x)$ is the final prediction,
- $h_m(x)$ is the weak classifier at iteration m,
- α_m is the weight assigned to the mth weak classifier based on its accuracy,
- M is the total number of boosting rounds.

Along with other classification models, this study used AdaBoost to test its performance in the selected dataset. Default decision stumps were set as weak learner, boost rounds, and learning rates were not manually tuned. It helped in consistent behavior with previous algorithms before training; Feature scaling was performed on this dataset. As this is a classification problem, we have assessed our models using classification metrics: accuracy, precision, recall, and F1-score. It is a very powerful technique that works great in reducing bias with simple models and also provides accuracy with moderately noisy data. In addition, it can be sensitive to outliers and lead to overfitting if the number of weak learners is not properly controlled. Bearing all that in mind, however, the AdaBoost algorithm does provide a solid baseline for performance and valuable information when comparing to more complex classifiers covered in later sections.

Although standard methods for medical classifications exist based on a number of various types of classifiers combined together to obtain greater overall accuracy than possible from the low-power (weak) classifier, ensemble approaches provide a widely used compromise between the accuracy obtained and the generalization ability. AdaBoost algorithm, where base classifiers are trained, the impact of misclassifications is increased when the final classification is performed, is the most commonly encountered algorithm in clinical decision support systems to improve performance by correcting the mistakes of the previous ones in the iterative training process. For example, the AdaBoost algorithm successfully identified features related to disease in a study [86] of prediction of chronic kidney disease, and the results are critical for early intervention. In a different study [87] examining heart disease, a combination of AdaBoost with K-fold cross-validation enhanced the generalizability of the model used, resulting in a higher accuracy of predicting the probability of patient survival. An example of a strong predictive classifier model [88] from their paper's experiments is the AdaBoost-based ensemble method achieved high classification performance and demonstrated how symmetries and interactions among different clinical variables dictate stroke risk. These studies, in addition to accuracy on model performance, highlight the value of AdaBoost algorithms in creating measurable positive impact on clinical diagnostic systems such as settlement, generalization ability, and possibly in early diagnosis as well.

3.3.1.8 Random Forest Algorithm

RF (Random Forest) [89] is a popular algorithm that is widely used in sensitive and large-scale supervised machine learning applications because of its high accuracy and robustness for classification and regression problems. Ensemble methods are another approach for classification, random forests or bagging follows an ensemble where during training, we generate a large number of decision trees, and on prediction we take the majority (voting based) of the classes for input instance. Random Forest is a super version of Decision Tree as its works on the concept of minimizing overfitting and variance with the help of randomness that comes into play during learning. Specifically,

in the forest, the trees are trained on a random subset of the training data (this step is known as bootstrapping or bagging when relating it to one particular tree) and at every split only a subset of the features are used rather than all the features. If you recall from earlier variations, the main reason for combining random data sampling with random feature selection was to allow the model to generalize better and therefore be a more robust model. The combined prediction of the ensemble model of T trees can be written as:

$$H(x) = \text{mode}\{h_t(x)\}_{t=1}^{T} \tag{3.8}$$

where:

- $H(x)$ is the final prediction for input x,
- $h_t(x)$ is the prediction of the tth individual decision tree,
- T is the total number of trees in the forest.

RF is one of the classification models presented in this research that was implemented to evaluate performance on the selected dataset. Random Forests are insensitive to the scale of features so no normalization was performed during the preprocessing step. We trained our model with the default parameters: we did not tune the number of trees, nor the individual tree depth. The performance of classifiers was evaluated using standard classification metrics of accuracy, precision, recall, and F1-score. Random Forest is known for its good performance combined with its ability to control overfitting, making it powerful for noisy datasets. It is also capable of using both numerical and categorical features, which is often not the case for traditional machine learning algorithms, and attach feature importance scores to make the classifier interpretable in terms of the features. Although it can be computationally expensive as compared to simpler models, Random Forest is a stable and effective model and thus will serve as a baseline to compare against for more complex classifiers in upcoming sections.

Medical classification problems often require algorithms which have both accuracy and generalisation capability, especially in cases where variables are highly interacting and the data is not well-distributed. However, up to now, it has been successfully used in clinical diagnosis systems due to its ensemble structure formed by learning a large number of decision trees with random subsamples and attributes: Random Forest has proven to be one of the effective algorithms. For example, a study [90] designed for predicting heart conditions based on a variety of different risk factors gave high levels of accuracy with a Random Forest model by evaluating the risk factors collectively and provided a good decision support mechanism for early diagnosis. Another study [91] showed that when the disease data was distributed unevenly, the algorithm discriminated between classes successfully, and depending on which of the classes was the minority class, the algorithm could identify more than some examples of members of that particular class. Similar to another example [92] for predicting amongst normal and atypical cases for diabetes, Random Forest was able to both accurately model complex relationships while preserving model stability over time with classification accuracy. The Random Forest algorithm also provides multi-parametric advantages, including high accuracy, stability in unstable data structures, interpretation of variable importance, direct contribution to clinical decision processes, etc., according to these studies.

3.3.1.9 Multilayer Perceptron Algorithm

MLP (The Multilayer Perceptron) [93] is one of the most widely used primarily for classification and regression tasks on complex, high-dimensional data. A feedforward artificial neural network that consists of an input layer, a hidden layer, or multiple hidden layers and an output layer. In a fully connected neural network, a neuron (or node) in one layer is connected to all the neurons present in the next layer, with weights assigned to these connections. MLP learns by adjusting its weights using an algorithm called backpropagation, which gradually propagates the error backwards through the network in order to minimize the difference between the predicted and actual output. To compute its output, each neuron makes a weighted summation of its inputs that is

passed through a nonlinear activation function such as *sigmoid*, *ReLU*, or *tanh*. The output α for a given neuron can be written as:

$$a = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{3.9}$$

where:

- x_i are the input features,
- w_i are the corresponding weights,
- b is the bias term,
- f is the activation function (e.g., *sigmoid*, *ReLU*),
- α_m is the neuron's output.

Another classification model used in the study to apply on the filtered dataset is Multilayer Perceptron (MLP). The MLP is sensitive to the scale of input features so that all features were•normalized during preprocessing to ensure stable training. Whatever the hyper parameters (layer count, neuron count, learning rate), it is set to the pre-defined default (did not try to optimize the hyper parameters). For evaluation on the test set, we calculated the standard classification metrics such as accuracy, precision, recall, and the weighted average F1-score. MLP can capture nonlinear dependencies in data with unknown complexities, thanks to its multiple hidden layers. But it usually demands more compute and is at risk of overfitting, especially for small datasets. However, similar to tree-based models, it does not have categorical variable•handling nor does it deliver inbuilt feature importance scores. Although its limitations also lie with MLP's capabilities to learn simply nonlinear decision boundaries, it provides a strong baseline when utilizing nonlinear functions, such as deep learning model to compare with simple classifiers in the•later sections.

However, it is essential to be able to model complex nonlinear relationships for instance where different clusters of symptoms appear in a disease. Within this frame, the Multilayer Perceptron (MLP) algorithm has risen to a prominent level with its multilayer structure and significant learning capability among the techniques used in clinical diagnosis systems. The study [94] on the prediction of heart diseases used the MLP neural network that was able to learn complex relationships between clinical variables with a high accuracy in the discrimination between patients and healthy individuals. As a result of the improvements made with the MLP architecture [95], the ability to diagnose thyroid diseases early was increased, and the contributing attributes were more effective in the processing of the features in another system, which is developed for the prediction of thyroid diseases. Moreover, MLP algorithms offered successful results by preserving both overall accuracy and sensitivity in chronic kidney disease prediction studies [96] performed with unbalanced data sets due to their flexible structure that can learn the samples belonging to the minority class. Such studies confirm the MLP as a promising technique in medical classification problems due to not only its higher performance but also it has the ability to learn nonlinear patterns, high generalization capability, and fidelity to diverse clinical scenarios.

CHAPTER 4

EXPERIMENTAL RESULTS

All the experimental results were obtained on a machine with 2.6 GHz Intel Core i7 CPU and 16 GB of RAM The proposed methods were implemented using MATLAB software and Python. 70% of the dataset was used for training while the remaining 30% was kept for testing. For performance measurement of the machine learning models, we present precision, recall, F1 score, and accuracy as shown in Table 4.1. In the subsequent subsections, we will report the capability of the machine learning approaches on all characteristics, statistical feature selection approaches and metaheuristic feature selection approaches individually.

Table 4.1 Performance metrics of the machine learning methods

Measure	**Formula**
Precision (Prec)	$\frac{TP}{TP+FP}$
Recall (Rec)	$\frac{TP}{TP+FN}$
F1 Score	2 x $\frac{PrecisionxRecall}{Precision+Recall}$
Accuracy (Acc)	$\frac{TP+TN}{TP+TN+FP+FN}$

4.1 Results of the machine learning classifiers without feature selection

In this part, we discuss the performance of the machine learning methods based on precision, recall, and F1-score metrics when all 38 features are used. Table 4.2 presents the precision, recall, and F1-score values of the machine learning classifiers. The precision values of the methods range from 0.47 to 0.87, with the best results obtained

using the RF. The recall values of the methods range from 0.46 to 0.86, while the highest recall value (0.86) is achieved with the RF. The F1-score values of the methods range from 0.43 to 0.86, whereas SVM has the lowest F1-score of 0.43 and the RF has the highest F1-score of 0.86. Furthermore, the accuracy values range from 61.46% to 86.46%. The LR reaches an accuracy of 85.42%, NB presents an accuracy of 61.46%, accuracy of the DT is 66.67%, the KNN reaches an accuracy of 48.96%, SVM presents an accuracy of 45.83%, the GB and AB achieves, and accuracy of 85.42%, RF presents an accuracy of 86.46%, accuracy of the MLP is 64.58%. The best accuracy (86.46%) is obtained when the RF is used. Finally, confusion matrices for each machine learning classifier are presented in Figure 4.1 without applying any feature selection method. As can be seen in this figure, the RF only misclassifies 5 out of 44 gallstone disease prediction cases, and it misclassifies 8 out of 52 non-gallstone disease prediction cases.

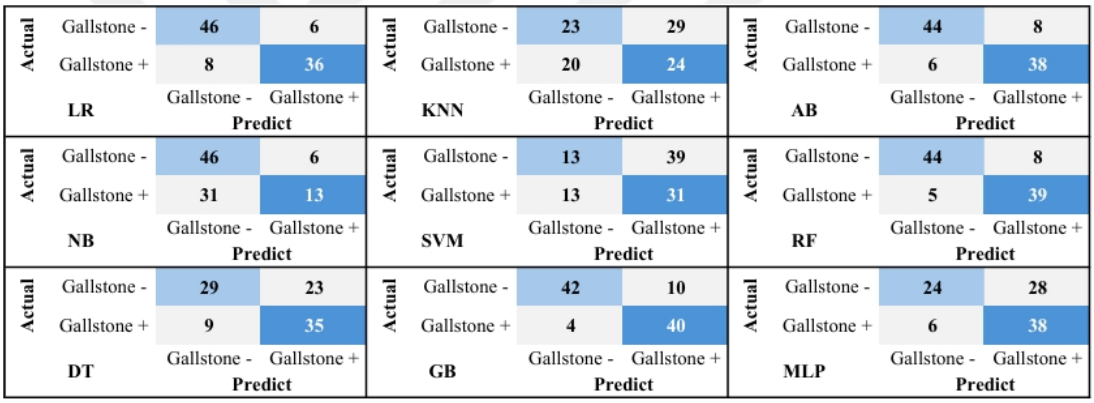

Figure 4.1 Confusion matrices for each machine learning classifier without feature selection.

Table 4.2 Performance of the machine learning classifiers without feature selection

	Prec	Rec	F1-score	Acc (%)
LR	0.85	0.85	0.85	85.42
NB	0.64	0.61	0.58	61.46
DT	0.69	0.67	0.66	66.67
KNN	0.50	0.49	0.49	48.96
SVM	0.47	0.46	0.43	45.83
GB	0.86	0.85	0.85	85.42
AB	0.86	0.85	0.85	85.42
RF	0.87	0.86	0.86	86.46
MLP	0.70	0.65	0.63	64.58

4.2 Results of the machine learning classifiers using statistical feature selection methods

In this part, we discuss the performance of the machine learning methods shown in Table 4.3 when the statistical feature selection methods that are MRMR, Chi^2, Anova, and Kruskal-Wallis are used. The first, we evaluate results of the MRMR feature selection method.

The MRMR selects 16 features among 38 features that are gender, comorbidity, CAD, hypothyroidism, hyperlipidemia, DM, ECF/TBW, LM, VFR, HFA, HDL, AST, ALP, CRP, hemoglobin, vitamin D. The accuracy results of the machine learning classifiers are between 0.57 and 0.87, and the best result is achieved when the RF is used. The precision values of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.79, 0.67, 0.73, 0.67, 0.72, 0.86, 0.85, 0.88 and 0.80, respectively. While the best result is obtained with the RF, the NB and the KNN have the lowest performance. The recall values of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.79, 0.57, 0.73, 0.66, 0.70, 0.86, 0.84, 0.88 and 0.74, respectively. While the best result is obtained with the RF, the NB has the lowest result. Furthermore, we evaluate F1-score results that range from 0.46 to 0.88. F1-score of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.79, 0.46, 0.73, 0.66, 0.70, 0.86, 0.84, 0.88 and 0.73, respectively. Finally, confusion matrices for each machine learning classifier are presented in Figure 4.2 when MRMR is only used. As can be seen in this figure, the RF only misclassifies 5 out of 44 gallstone disease prediction cases, and it misclassifies 7 out of 52 non-gallstone disease prediction cases.

Table 4.3 Results of the machine learning methods when statistical feature selection methods are used.

ML Methods	Acc (%)	Prec	Rec	F1	Acc (%)	Prec	Rec	F1	Acc (%)	Prec	Rec	F1	Acc (%)	Prec	Rec	F1
LR	79.17	0.79	0.79	0.79	81.25	0.81	0.81	0.81	86.46	0.86	0.86	0.86	83.33	0.83	0.83	0.83
NB	57.29	0.67	0.57	0.46	60.42	0.71	0.60	0.52	67.71	0.68	0.68	0.67	64.58	0.75	0.65	0.59
DT	72.92	0.73	0.73	0.73	72.92	0.73	0.73	0.73	68.75	0.70	0.69	0.69	78.12	0.78	0.78	0.78
KNN	65.62	0.67	0.66	0.66	62.50	0.64	0.62	0.62	64.58	0.65	0.65	0.65	66.67	0.67	0.67	0.67
SVM	69.79	0.72	0.70	0.70	64.58	0.65	0.65	0.65	75.00	0.75	0.75	0.75	61.46	0.62	0.61	0.61
GB	86.46	0.86	0.86	0.86	82.29	0.83	0.82	0.82	85.42	0.86	0.85	0.85	83.33	0.84	0.83	0.83
AB	84.38	0.85	0.84	0.84	84.38	0.85	0.84	0.84	87.50	0.88	0.88	0.87	86.46	0.87	0.86	0.86
RF	87.50	0.88	0.88	0.88	87.50	0.88	0.88	0.88	85.42	0.85	0.85	0.85	87.50	0.88	0.88	0.88
MLP	73.96	0.80	0.74	0.73	80.21	0.80	0.80	0.80	77.08	0.79	0.77	0.77	62.50	0.67	0.62	0.61

The Chi^2 selects 23 features among 38 features that are gender, CAD, hyperlipidemia, DM, ECW, ICW, ECF/TBW, TBFR, LM, protein, VFR, BM, TFC, VFA, VMA, HFA, TC, HDL, AST, creatinine, CRP, hemoglobin, vitamin D. The accuracy results of the

machine learning classifiers are between 0.60 and 0.87, and the best result is achieved when the RF is used. The precision values of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.81, 0.71, 0.73, 0.64, 0.65, 0.83, 0.85, 0.88 and 0.80, respectively. While the best result is obtained with the RF, the KNN has the lowest performance. The recall values of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.81, 0.60, 0.73, 0.62, 0.65, 0.82, 0.84, 0.88 and 0.80, respectively. While the best result is obtained with the RF, the NB has the lowest result. Furthermore, we evaluate F1-score results that range from 0.52 to 0.88. F1-score of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.81, 0.52, 0.73, 0.62, 0.65, 0.82, 0.84, 0.88 and 0.80, respectively. Finally, confusion matrices for each machine learning classifier are presented in Figure 4.3 when Chi^2 is only used. As can be seen in this figure, the RF only misclassifies 3 out of 44 gallstone disease prediction cases, and it misclassifies 9 out of 52 non-gallstone disease prediction cases.

The ANOVA selects 9 features among 38 features that are ECW, ECF/TBW, TBFR, LM, BM, TFC, CRP, hemoglobin, vitamin D. The accuracy results of the machine learning classifiers are between 0.64 and 0.87, and the best result is achieved when the AB is used. The precision values of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.86, 0.68, 0.70, 0.65, 0.75, 0.86, 0.88, 0.85 and 0.79, respectively. While the best result is obtained with the AB, the KNN has the lowest performance. The recall values of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.86, 0.68, 0.69, 0.65, 0.75, 0.85, 0.88, 0.85 and 0.77, respectively. While the best result is obtained with the AB, the KNN has the lowest result. Furthermore, we evaluate F1-score results that range

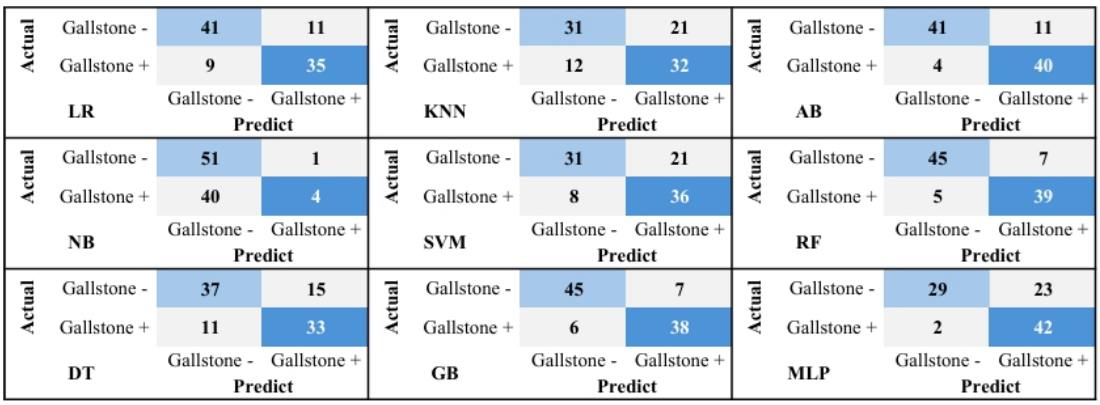

Figure 4.2 Confusion matrices for each machine learning classifier when MRMR is used.

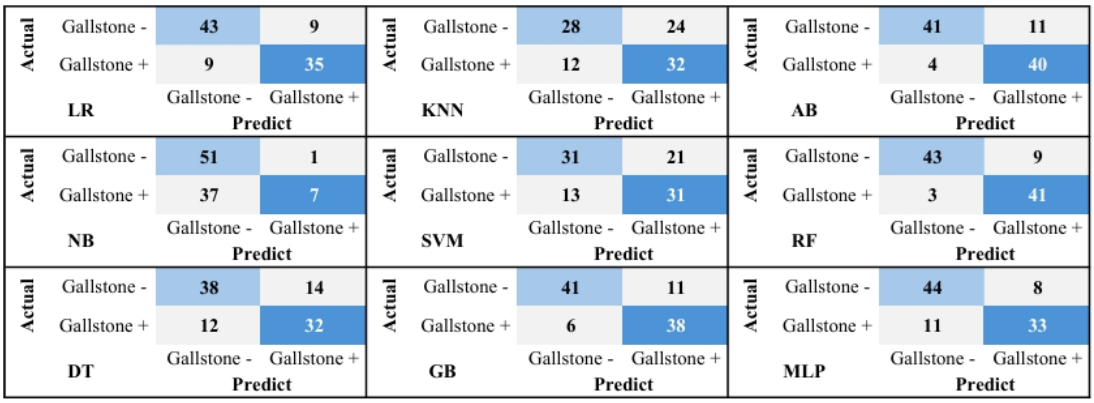

Figure 4.3 Confusion matrices for each machine learning classifier when Chi² is used.

from 0.65 to 0.87. F1-score of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.86, 0.67, 0.69, 0.65, 0.75, 0.85, 0.87, 0.85 and 0.77, respectively. Finally, confusion matrices for each machine learning classifier are presented in Figure 4.4 when ANOVA is only used. As can be seen in this figure, the AB only misclassifies 7 out of 44 gallstone disease prediction cases, and it misclassifies 5 out of 52 non-gallstone disease prediction cases.

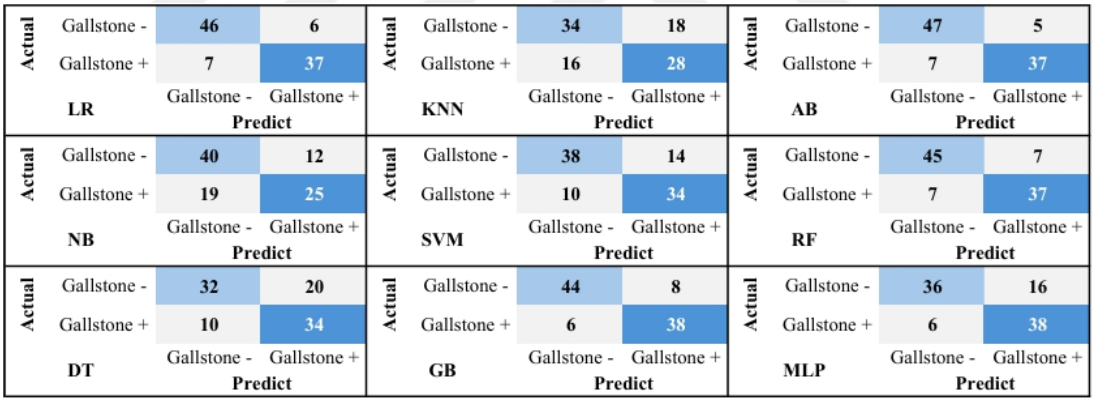

Figure 4.4 Confusion matrices for each machine learning classifier when ANOVA is used.

The Kruskal-Wallis selects 33 features among 38 features that are age, gender, CAD, hypothyroidism, hyperlipidemia, DM, height, weight, BMI, TBW, ECW, ICW, ECF/TBW, TBFR, LM, protein, BM, MM, TFC, VFA, VMA, HFA, Glucose, LDL, HDL, AST, ALT, ALP, creatinine, GFR, CRP, hemoglobin, vitamin D. The accuracy results of the machine learning classifiers are between 0.61 and 0.87, and the best result is achieved when the RF is used. The precision values of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.83, 0.75, 0.78, 0.67, 0.62, 0.84, 0.87, 0.88 and 0.67, respectively. While the best result is obtained with the RF, the SVM has the

lowest performance. The recall values of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.83, 0.65, 0.78, 0.67, 0.61, 0.83, 0.86, 0.88 and 0.62, respectively. While the best result is obtained with the RF, the SVM has the lowest result. Furthermore, we evaluate F1-score results that range from 0.61 to 0.88. F1-score of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.83, 0.59, 0.78, 0.67, 0.61, 0.83, 0.86, 0.88 and 0.61, respectively. Finally, confusion matrices for each machine learning classifier are presented in Figure 4.5 when Kruskal-Wallis is only used. As can be seen in this figure, the RF only misclassifies 3 out of 44 gallstone disease prediction cases, and it misclassifies 9 out of 52 non-gallstone disease prediction cases.

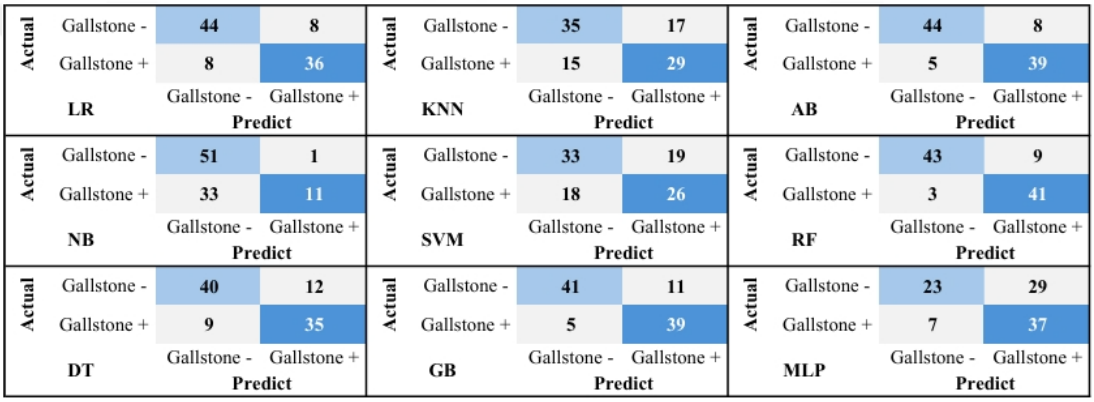

Figure 4.5 Confusion matrices for each machine learning classifier when Kruskal-Wallis is used.

4.3 Results of the machine learning classifiers using metaheuristic feature selection methods

In this section, we discuss the performance of the machine learning classifiers based on precision, recall, F1-score, and accuracy shown in Table 4.4 when the GWO, PSO, WOA, and HHO feature selection methods are used. First we analyze the performance of the machine learning methods when the GWO is used in the feature selection process.

The GWO selects 27 features among 38, which are age, gender, comorbidity, CAD, hyperlipidemia, height, weight, BMI, ECW, ICW, ECF/TBW, TBFR, LM, protein, BM, MM, obesity, TFC, VFA, VMA, HFA, AST, creatinine, GFR, CRP, hemoglobin, vitamin D. Table 4.4 presents precision, recall, F1-score, and accuracy results of the machine learning methods when the GWO is used in the feature selection process. The

Table 4.4 Results of the machine learning methods when the metaheuristic algorithms are used in the feature selection process

	GWO				PSO				WOA				HHO			
ML Methods	**Acc**	**Prec**	**Rec**	**F1**	**Acc**	**Prec**	**Rec**	**F1**	**Acc**	**Prec**	**Rec**	**F1**	**Acc**	**Prec**	**Rec**	**F1**
LR	81.25	0.81	0.81	0.81	81.25	0.81	0.81	0.81	79.17	0.79	0.79	0.79	86.46	0.86	0.86	0.86
NB	59.38	0.70	0.59	0.50	59.38	0.77	0.59	0.49	59.38	0.70	0.59	0.50	60.42	0.71	0.60	0.52
DT	78.12	0.79	0.78	0.78	76.04	0.77	0.76	0.76	80.21	0.81	0.80	0.80	79.17	0.80	0.79	0.79
KNN	67.71	0.68	0.68	0.68	62.50	0.63	0.62	0.63	64.58	0.65	0.65	0.65	62.50	0.63	0.62	0.63
SVM	71.88	0.72	0.72	0.72	67.71	0.68	0.68	0.68	69.79	0.70	0.70	0.70	73.96	0.75	0.74	0.74
GB	85.42	0.85	0.85	0.85	88.54	0.89	0.89	0.89	84.38	0.84	0.84	0.84	85.42	0.86	0.85	0.85
AB	87.50	0.88	0.88	0.88	85.42	0.86	0.85	0.85	88.54	0.89	0.89	0.89	88.54	0.89	0.89	0.89
RF	89.58	0.90	0.90	0.90	89.58	0.90	0.90	0.90	90.62	0.91	0.91	0.91	83.33	0.83	0.83	0.83
MLP	78.12	0.79	0.78	0.78	70.83	0.72	0.71	0.70	65.62	0.78	0.66	0.63	69.79	0.78	0.70	0.68

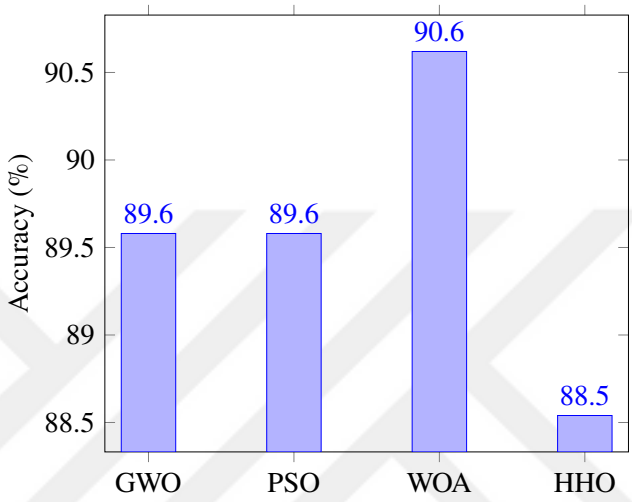

Figure 4.6 Accuracy results of the machine learning methods using metaheuristic feature selection methods

accuracy results of the machine learning classifiers are between 0.59 and 0.89, and the best result is achieved when the RF is used. The precision values of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.81, 0.70, 0.79, 0.68, 0.72, 0.85, 0.88, 0.90 and 0.79, respectively. While the best result is obtained with the RF, the KNN has the lowest performance. The recall values of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.81, 0.59, 0.78, 0.68, 0.72, 0.85, 0.88, 0.90 and 0.78, respectively. While the best result is obtained with the RF, the NB has the lowest result. Furthermore, we evaluate F1-score results that range from 0.50 to 0.90. F1-score of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.81, 0.50, 0.78, 0.68, 0.72, 0.85, 0.88, 0.90 and 0.78, respectively. Finally, confusion matrices for each machine learning classifier are presented in Figure 4.7 when GWO is only used. As can be seen in this figure, the RF only misclassifies 5 out of 44 gallstone disease prediction cases, and it misclassifies 5 out of 52 non-gallstone disease prediction cases.

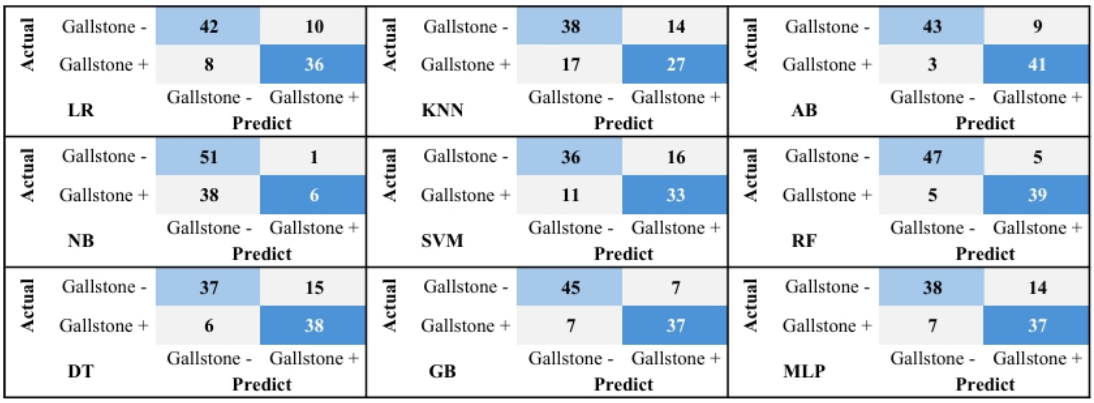

Figure 4.7 Confusion matrices for each machine learning classifier when GWO is used.

The PSO selects 18 features among 38, which are comorbidity, hyperlipidemia, DM, ECW, ICW, ECF/TBW, TBFR, LM, BM, obesity, TFC, VFA, VMA, glucose, ALP, CRP, hemoglobin, vitamin D. Table 4.4 presents precision, recall, F1-score, and accuracy results of the machine learning methods when the PSO is used in the feature selection process. The accuracy results of the machine learning classifiers are between 0.59 and 0.89, and the best result is achieved when the RF is used. The precision values of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.81, 0.77, 0.77, 0.63, 0.68, 0.89, 0.86, 0.90 and 0.72, respectively. While the best result is obtained with the RF, the KNN has the lowest performance. The recall values of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.81, 0.59, 0.76, 0.62, 0.68, 0.89, 0.85, 0.90 and 0.71, respectively. While the best result is obtained with the RF, the NB has the lowest result. Furthermore, we evaluate F1-score results that range from 0.49 to 0.90. F1-score of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.81, 0.49, 0.76, 0.63, 0.68, 0.89, 0.85, 0.90 and 0.70, respectively. Finally, confusion matrices for each machine learning classifier are presented in Figure 4.8 when PSO is only used. As can be seen in this figure, the RF only misclassifies 4 out of 44 gallstone disease prediction cases, and it misclassifies 6 out of 52 non-gallstone disease prediction cases.

The WOA selects 27 features among 38, which are age, comorbidity, hypothyroidism, hyperlipidemia, DM, height, weight, BMI, ECW, ICW, ECF/TBW, TBFR, LM, protein, VFR, BM, MM, obesity, TFC, VFA, VMA, HFA, AST, creatinine, CRP, hemoglobin, vitamin D. Table 4.4 presents precision, recall, F1-score, and accuracy results of the

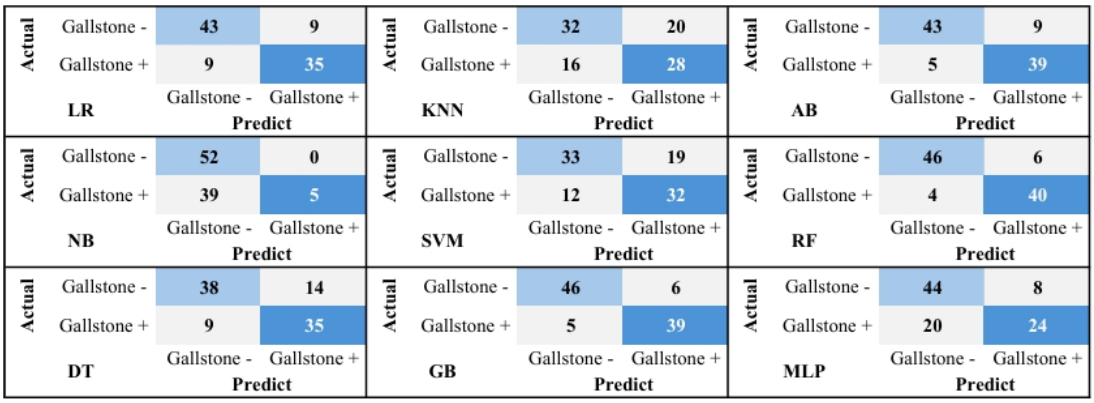

Figure 4.8 Confusion matrices for each machine learning classifier when PSO is used.

machine learning methods when the WOA is used in the feature selection process. The accuracy results of the machine learning classifiers are between 0.59 and 0.90, and the best result is achieved when the RF is used. The precision values of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.79, 0.70, 0.81, 0.65, 0.70, 0.84, 0.89, 0.91 and 0.78, respectively. While the best result is obtained with the RF, the KNN has the lowest performance. The recall values of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.79, 0.59, 0.80, 0.65, 0.70, 0.84, 0.89, 0.91 and 0.66, respectively. While the best result is obtained with the RF, the NB has the lowest result. Furthermore, we evaluate F1-score results that range from 0.50 to 0.91. F1-score of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.79, 0.50, 0.80, 0.65, 0.70, 0.84, 0.89, 0.91 and 0.63, respectively. Finally, confusion matrices for each machine learning classifier are presented in Figure 4.9 when WOA is only used. As can be seen in this figure, the RF only misclassifies 3 out of 44 gallstone disease prediction cases, and it misclassifies 6 out of 52 non-gallstone disease prediction cases.

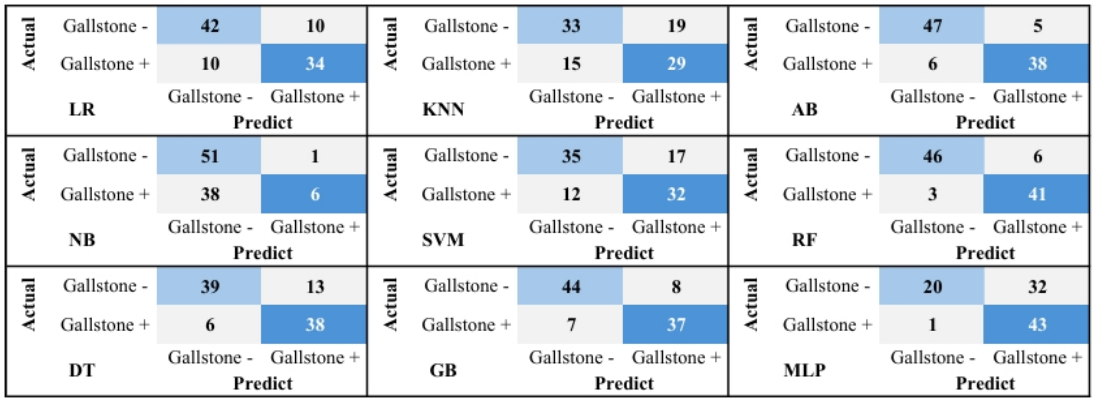

Figure 4.9 Confusion matrices for each machine learning classifier when WOA is used.

The HHO selects 24 features among 38, which are age, comorbidity, CAD, hypothyroidism, hyperlipidemia, DM, BMI, TBW, ECW, ICW, ECF/TBW, TBFR, LM, BM, obesity, TFC, VFA, AST, ALT, ALP, creatinine, CRP, hemoglobin, vitamin D. Table 4.4 presents precision, recall, F1-score, and accuracy results of the machine learning methods when the HHO is used in the feature selection process. The accuracy results of the machine learning classifiers are between 0.60 and 0.88, and the best result is achieved when the AB is used. The precision values of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.86, 0.71, 0.80, 0.63, 0.75, 0.86, 0.89, 0.83 and 0.78, respectively. While the best result is obtained with the AB, the KNN has the lowest performance. The recall values of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.86, 0.60, 0.79, 0.62, 0.74, 0.85, 0.89, 0.83 and 0.70, respectively. While the best result is obtained with the AB, the NB has the lowest result. Furthermore, we evaluate F1-score results that range from 0.52 to 0.89. F1-score of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.86, 0.52, 0.79, 0.63, 0.74, 0.85, 0.89, 0.83 and 0.68, respectively. Finally, confusion matrices for each machine learning classifier are presented in Figure 4.10 when HHO is only used. As can be seen in this figure, the AB only misclassifies 3 out of 44 gallstone disease prediction cases, and it misclassifies 8 out of 52 non-gallstone disease prediction cases.

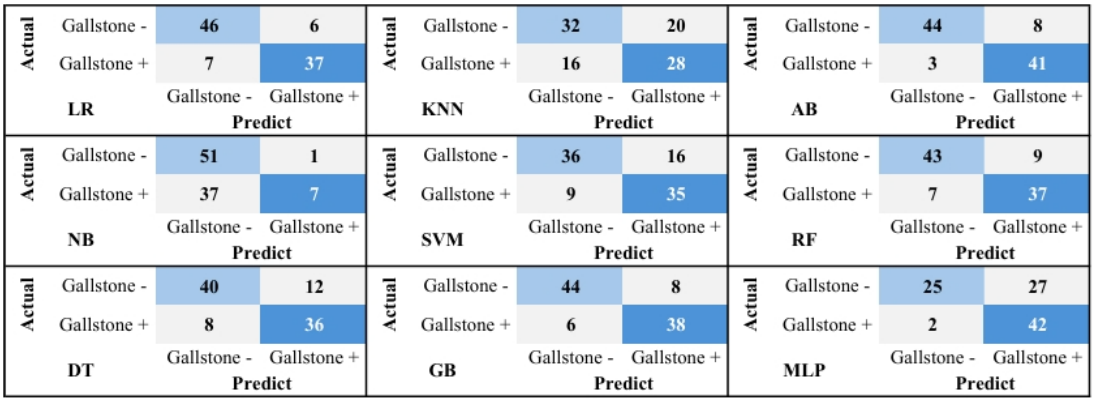

Figure 4.10 Confusion matrices for each machine learning classifier when HHO is used.

4.4 Results of Proposed Ensemble Feature Selection Method

The proposed ensemble feature selection method selects 11 features among 38, which are hyperlipidemia, ECW, ECF/TBW, TBFR, LM, BM, TFC, VFA, CRP, hemoglobin, and vitamin D. Table 4.5 presents precision, recall, F1-score, and accuracy results of

the machine learning methods when the ensemble feature selection method is used. The accuracy results of the machine learning classifiers are between 0.57 and 0.90, and the best result is achieved when the RF is used. The precision values of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.83, 0.76, 0.77, 0.68, 0.76, 0.89, 0.90, 0.91 and 0.81, respectively. While the best result is obtained with the RF, the KNN has the lowest performance. The recall values of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.83, 0.57, 0.76, 0.68, 0.76, 0.89, 0.90, 0.91 and 0.80, respectively. While the best result is obtained with the RF, the NB has the lowest result. Furthermore, we evaluate F1-score results that range from 0.45 to 0.91. F1-score of the LR, NB, DT, KNN, SVM, GB, AB, RF, and MLP are 0.83, 0.45, 0.76, 0.68, 0.76, 0.89, 0.90, 0.91 and 0.80, respectively. Finally, confusion matrices for each machine learning classifier are presented in Figure 4.11 when the proposed ensemble feature selection method is used. As can be seen in this figure, the RF only misclassifies 4 out of 44 gallstone disease prediction cases, and it misclassifies 5 out of 52 non-gallstone disease prediction cases.

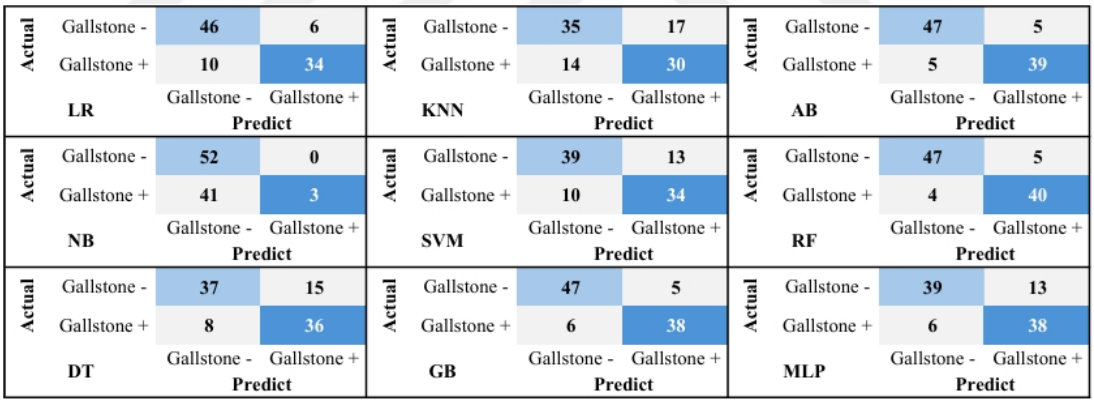

Figure 4.11 Confusion matrices for each machine learning classifier when the proposed ensemble feature selection method is used.

Figure 4.6 gives comparison of machine learning methods with respect to different feature selection strategies. The results are shown based on the best performing feature selection method considering accuracy performance measure. As can be seen in Figure 4.6, when all 38 features are used, the RF achieves the best accuracy of 86%. When the results of the statistical feature selection methods are evaluated, the Anova gives better results and selects 9 features. The Adaboost with Anova achieves the best result with an accuracy of 88%. When the results of the metaheuristic feature selection methods

Table 4.5 Results of the machine learning methods when the proposed feature selection method is used.

	Prec	**Rec**	**F1-score**	**Acc (%)**
LR	0.83	0.83	0.83	83.33
NB	0.76	0.57	0.45	57.29
DT	0.77	0.76	0.76	76.04
KNN	0.68	0.68	0.68	67.71
SVM	0.76	0.76	0.76	76.04
GB	0.89	0.89	0.89	88.54
AB	0.90	0.90	0.90	89.58
RF	0.91	0.91	0.91	90.62
MLP	0.81	0.80	0.80	80.21

are evaluated, the WOA gives better results and selects 27 features. The RF with WOA achieves the best result with an accuracy of 91%. Finally, when the results of the proposed method are evaluated, the proposed ensemble feature selection method selects 11 features. The RF with the proposed ensemble feature selection method achieves the best result with an accuracy of 91%. When comparing the proposed method with the WOA, the proposed method reaches the same accuracy with the WOA using relatively a few number of features. This underlines the efficiency of the proposed method, and the proposed feature selection method significantly reduces classification complexity.

4.5 Comparison with the-state-of-the-art gallstone detection methods

Several studies have been published for the diagnosis of gallstones using artificial intelligence approaches. These studies employ artificial intelligence-based techniques to analyze both textual and imaging data obtained from patients, aiming to improve diagnostic accuracy and reduce the reliance on manual interpretation by medical professionals. Imaging-based approaches primarily focus on ultrasound, CT, or MRI scans, utilizing deep learning and computer vision techniques to detect gallstones with high precision. On the other hand, textual data-driven studies analyze patient records, clinical notes, and laboratory findings using machine learning algorithms to extract

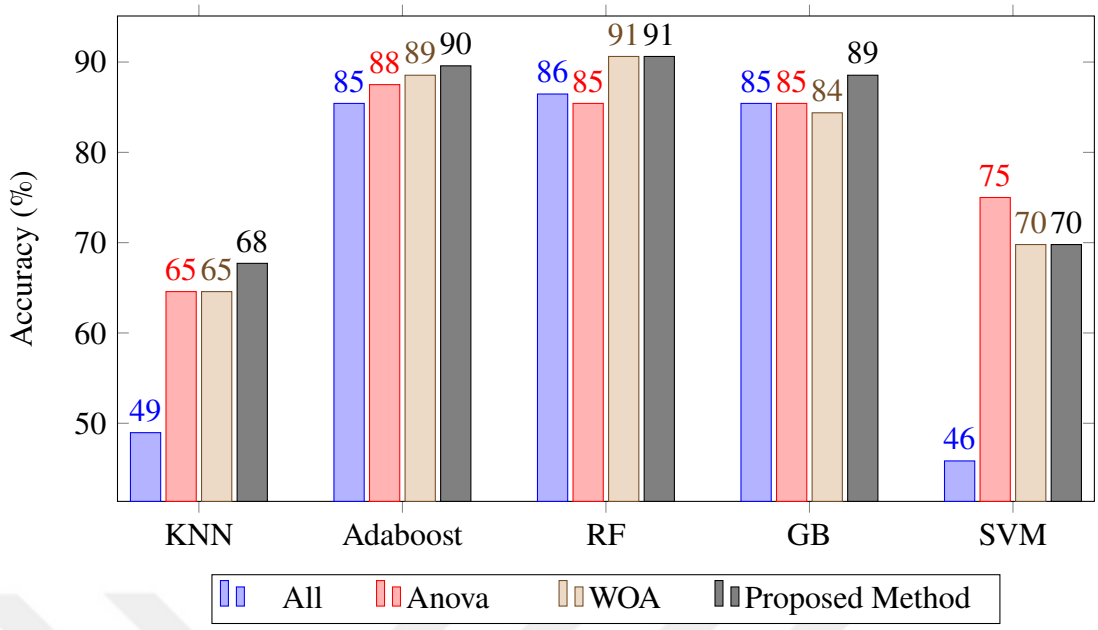

Figure 4.12 Accuracy results of the machine learning methods when all features (without feature selection) are used, and ANOVA, WOA as well as the proposed ensemble feature selection method are applied in the feature selection process

meaningful patterns indicative of gallbladder disease. In this section, we compare the proposed method with existing studies that detect gallstones using datasets that do not include images.

Salem et al. [6] proposed based on a machine learning and deep learning study for the diagnosis of gallstones from metabolomics and clinical data. Their dataset includes 218 features belong to 100 patients with gallstones and 118 patients with healthy control. They achieved an accuracy on 82% when the CNN with 50 metabolites and 5 clinical features was used. Liew et al. [8] compared the artificial neural network model with the logistic regression model to predict gallstone disease in obese individuals. When using clinical and laboratory data collected from 117 patients, their proposed ANN and logistic regression method achieved an accuracy of 97.14% and 88.2%, respectively. Himi et al. [18] presented a system that integrates smartwatches and machine learning to predict common diseases. Their study aimed to predict diseases such as hypertension, diabetes, kidney failure, and gallstones using a smartwatch equipped with sensors that measure body statistics and machine learning algorithms. They collected the data from a prototype smartwatch including 11 sensors. Their dataset including 20 features coming from 260 patients. They applied various types of machine learning methods

and their experimental results presented that the LR method achieved an accuracy of 96.67%.

Recently, Esen et al. [26] proposed a machine learning-based study for the early prediction of gallstone disease. They used various machine learning methods. Their dataset contained 38 features, including bioimpedance and laboratory information collected from 319 patients, of whom 161 had gallstones and 158 were healthy controls. Their experimental results showed that the gradient boosting method achieved the highest accuracy (85.42%) for predicting gallstones. In this study, we improve their approach by introducing ensemble feature selection method. Our experimental results show that selecting informative features significantly reduces the total number of false negatives and false positives, and the proposed method achieves 90.62% accuracy on the dataset used by Esen et al. [26]. To sum up, when compared to the state-of-the-art gallstone detection methods, the proposed method detects gallstones with higher accuracy using relatively few number of features. This highlights that using an optimal set of features improves overall accuracy.

CHAPTER 5

CONCLUSION AND FUTURE STUDIES

5.1 Conclusion Remarks

Gallstone disease is a common global health problem with the potential for significant morbidity if diagnosis and management are delayed. Currently, there are limitations for imaging techniques widely used for diagnosis, especially for ultrasonography, CT, and MRI. Ultrasound is highly operator-dependent, and diagnostic discrepancies may occur as the interpretation is subjective and highly dependent on the experience of the practitioner. In addition, these imaging modalities are costly, needing expensive equipment and skilled personnel. This complicates getting a diagnosis, particularly for those in health care-strapped parts of the world.

In this study, the authors propose the development of a machine learning-driven diagnostic model that can predict gallstone disease based solely on laboratory findings and bioimpedance data, as a solution to the drawbacks mentioned above. Therefore, the primary aim of the work is to establish a model that can provide extremely precise forecast using publically available and inexpensive data without the necessity of imaging techniques. A dataset with 38 features comprising the demographic, biochemical, and bioimpedance information of 319 patients who had been treated at the Ankara VM Medical Park Hospital was used for this objective.

Due to the high dimensionality of the dataset and the potential for multicollinearity, we begin with an efficient feature selection process. Overall, four statistical feature selection algorithms (MRMR (Minimum Redundancy Maximum Relevance), Chi^2 (Chi-Square Test), ANOVA (Analysis of Variance) and Kruskal-Wallis) and four current metaheuristics (Grey Wolf Optimization (GWO), Whale Optimization Algorithm (WOA), Particle Swarm Optimization (PSO), Harris Hawk Optimization (HHO)) were applied for this process. Thirteen algorithms achieved accuracy under different proposed discrete ensemble feature selection methods, and methods of the visualization of useful features to a list of 11 features that contribute significantly to

classifiers that were selected by at least 7 of these 8 algorithms. Hence, we seek to have a simpler, interpretable, and faster model which is working on meaningful and influencing variables.

A total of nine different machine learning algorithms were employed for classification using these 11 selected features and each algorithm was assessed in terms of key performance measures including accuracy, sensitivity, specificity, and F1 score (Logistic Regression, Naive Bayes, Support Vector Machines, Decision Trees, K-Nearest Neighbor, Gradient Boosting, AdaBoost, Random Forest, and Multilayer Perceptron). Specifically, the Random Forest algorithm, applied in the proposed feature selection method, made a 90.62 percent correct prediction with 11 features and was found to be the most successful classifier of the analysis.

Not only does the model's success affect the accuracy rate, it also reduces classification complexity and provides high performance using fewer features, respectively, so that the algorithms can be trained in the fewer time. In this way, it is not just an academic success, it would also be a practical solution that can be incorporated into clinical workflows. Besides, working on raw data, the absence of missing data, and the balance of the sample group are among the factors that increase the reliability of the results.

The method provides in this study achieved better performance in terms of feature efficiency and accuracy rates compared to gallstone diagnosis approaches available in the literature. Especially, compared to Esen et al. [26] who had studied with the same dataset and only used the Gradient Boosting algorithm to achieve an accuracy rate of 85.42 percent, it is clearly seen that the great success of the feature selection made with the proposed method and the power of the model as Esen et al. [26] who worked on the same data set only created Gradient Boosting algorithm.

Such a diagnostic system, which is developed in this thesis, could be incorporated in to a clinical decision support system, and thus provide the clinician with the necessary information to gauge the need for additional imaging studies in a fast and cheap manner, and thus potentially meet the need in areas with poor access to advanced imaging instrumentation. When tested in different patient populations in the

upcoming studies, when validated with bigger multicenter databases, and when user-friendly interfaces are settled and supported by explainable artificial intelligence techniques, this approach will be used wider in the health area. This study is a very good case of AI-aided diagnostic systems, and suggests that such processes might be a part of the diagnosis of other diseases too.

5.2 Future Work

Although our proposed ensemble-based machine learning framework has shown strong promise at accurately diagnosing gallstone disease utilizing clinical and bioimpedance data, this is a broad field that allows for plenty of room in the improvement, validation, and practical implementation of the results. This study represents an encouraging first step towards the development of generic, low-cost, accurate, and accessible clinical decision support systems. However, future research needs a multidisciplinarity and iterative approach to enhance its clinical relevance, technical robustness, and user acceptance. Thus, there are critical directions such as increasing diversity in data, increasing the complexity of algorithms, always making interpretability and real-time accessible along with incorporation of the model in clinical workflows that need to be explored to transform this framework into a full-fledged deployable diagnostic machine which will be crucial in improving healthcare delivery and optimizing patient outcomes on a larger scale.

5.2.1 Dataset Generalization and Population Diversity

The key implication and follow up is for the sample to be extended and further diversified for validation. The model may not extrapolate well to larger populations since the data were collected at a fixed timepoint from a single hospital. Indeed, a key goal is to include multi-institutional datasets that capture a broad array of demographic characteristics, geographic areas, and health-care settings in future novice pathways research. Having access to such datasets would enable the model to account for differing patient behavior, socioeconomic status, and comorbidity profiles, thereby mitigating against biases and

improving the fairness of clinical predictions.

Furthermore, stratified performance analysis according to different subgroups (e.g., age, sex, BMI, comorbidities) would allow deeper insight into the reliability and diagnostic consistency of the model. It would also facilitate the development of models or calibration strategies specific to such a subgroup.

5.2.2 Incorporation of Multimodal and Longitudinal Data

Although this study based the model on laboratory and bioimpedance features, the inclusion of additional data modalities will enhance its future iterations. These might be genetic markers, microbiome compositions, electronic health record histories, radiological findings, dietary patterns, and patient-reported symptoms. This would allow models to encompass a richer representation of gallstone pathogenesis through the integration of such data sources.

Additionally, the longitudinal nature of the data provides unique potential for modeling disease progression. The introductive utilization of time-series analysis with RNNs (Recurrent Neural Networks), GRUs (Gated Recurrent Units), or transformer architectures will uncover the time-dependent machinery used for gallstone agravation. This would allow for early warning systems and patient-centric monitoring systems, greatly improving the use of the model for preventive care.

5.2.3 Data Quality, Missing Values, and Class Imbalance Handling

Handling missing data and class imbalance, frequently observed in real-world clinical datasets, is another aspect this work must address in the future. While the present dataset is complete and well-distributed, subsequent datasets might have variance in missing/complete data owing to several operational factors like human errors, differential data entry, or varied measurement protocols.

However, in order for the model to be robust under these conditions, data itself must

be missing (which could be treated through statistical imputation methods like mean, median, regression, or more advanced methods like KNN (K-Nearest Neighbor). By including these preprocessing steps, we can minimize information loss as well as improve model generalizability.

Likewise, future investigations will need to be mindful of potential class imbalance—this could be particularly relevant when rare subtype specimens or early-phase gallstones are the subjects of analysis. To mitigate skewed distributions, one can apply techniques like SMOTE (Synthetic Minority Over-sampling Technique), ADASYN (Adaptive Synthetic) sampling, or cost-sensitive learning approaches. These approaches help to prevent the model from becoming biased towards the larger class while still being able to detect the minority cases that are vital for medical diagnosis.

5.2.4 Enhanced Feature Engineering and Metaheuristic Exploration

The Ensemble feature selection method is one of the approaches proposed in this work which can be further explored in future works. Such a fixed cut-off (e.g., requires features to be selected by 7 out of 8 methods) does not leverage the vote information all methods provide; so instead of a fixed vote-use threshold, we can introduce dynamic weighting and voting mechanisms to adaptively refine feature subsets.

Furthermore, additional meta-heuristic algorithms like Genetic Algorithms, Ant Colony Optimization, Simulated Annealing, Cuckoo Search, etc., may also be tested on complex feature space as comparison with the proposed methods in these cases. More informative and compact feature sets could be the result of hybrid feature selection pipelines that include statistical filters, embedded methods, and metaheuristics.

Further, when running on large-scale unstructured datasets in the future, unsupervised deep representation learning such as variational autoencoders and self-supervised contrastive learning could also discover latent (unobserved) structures in the data.

Selected features can also be reviewed and validated with domain experts to ensure the clinical interpretability of the selected features, in parallel.

5.2.5 Advancement of Machine Learning Models and Optimization Techniques

All the algorithms performed in the study, ranging from ensemble classifiers to neural networks, but future work could apply deeper networks. CNNs (Convolutional Neural Networks), and GNNs (Graph Neural Networks) might also account for more intricate relations among features, whereas attention-based models could capture dependencies across a wider context.

Automatic hyper-parameter tuning methods like Grid Search, Random Search, Bayesian Optimization, or using advanced frameworks like Optuna (Open-Source Hyperparameter Optimization Framework), and AutoML (Automated Machine Learning) can be applied to boost the performance of the model further. They search the parameter space in a systematic manner, meaning that they may find configurations that yield significant performance improvements without causing overfitting.

Furthermore, ensemble learning strategies (e.g., stacking, weighted averaging, or dynamic classifier selection) can be used to combine the strengths of multiple models, particularly when dealing with heterogeneous or noisy clinical data.

5.2.6 Interpretability and Clinical Trust via Explainable AI

One of the most significant obstacles to clinical implementation of AI remains model interpretability. Therefore, future research effort should concentrate on the application of XAI (Explainable Artificial Intelligence) tools to understand how predictive modelling is achieved. SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-Agnostic Explanations), and Integrated Gradients are approaches that shed light on the input features that affect the predictions and the reasons for them, which can help clinicians interpret predictions.

Furthermore, case-based reasoning and visualization dashboards to depict model rationales could improve physician confidence and trust. In addition, providing risk factors alongside interpretable outputs enables the model to augment clinical reasoning as opposed to function as a black-box tool.

5.2.7 Deployment in Real-Time and Embedded Systems

Adapting the model for real-time operation is necessary to make it clinically actionable. This could include converting the trained model to a lighter, optimized version that is suitable for deployment on mobile devices, edge devices, or embedded systems through frameworks such as TensorFlow Lite or ONNX.

These real-time systems could be integrated into outpatient clinics, emergency departments, or remote monitoring applications, particularly in regions with scarce imaging capabilities. When appropriately calibrated and validated, the model could aid in rapid triage decision-making, minimize delays to diagnosis, and enable prompt therapies.

5.2.8 Integration into Clinical Decision Support Systems (CDSS)

This model will eventually be integrated into CDSS (Clinical Decision Support Systems). By integrating the diagnostic algorithm into HIS (Hospital Information Systems), EHR (Electronic Health Records), or mobile health applications, clinicians have access to automated data-driven recommendations as part of standard patient assessments.

User interface development in which the predictions, confidence levels, and supporting evidence can be visualized in e.g. real time can facilitate system use. Such interfaces need to be developed in conjunction with clinicians to ensure that they seamlessly fit into existing workflows and clinical practices.

5.2.9 Clinical Trials, Usability Testing, and Physician Engagement

Before the system can be widely deployed, it will have to undergo extensive testing in prospective clinical trials and piloting studies. They may assess diagnostic accuracy, time savings, physician satisfaction, and patient outcomes in real-world circumstances. They can also uncover ethical or logistical issues that were invisible during the development of the algorithm.

At the same time, usability testing with physicians, nurses, and administrators should be undertaken to discover barriers to adoption. The interface should rely on human-centered design principles so that it can be intuitive, actionable, and responsive to the needs of all stakeholders.

Standardized interviews, questionnaires, or observational methods might be used to understand how the model aids decision-making and what refinements would be needed for clinical acceptance.

5.2.10 Patient-Facing Applications and Remote Healthcare

Patient-centered applications may also be developed, in addition to those for clinician use. At the end, mobile apps that allow patients to enter key parameters and immediately receive personalized risk scores could be the tools to facilitate patients to manage their health. These types of tools may also offer significant value in rural, underdeveloped, or low-resource areas where access to diagnostic imaging is impeded.

Integration with wearable devices and IoT (Internet of Things) platforms may allow more passive, continuous health monitoring. Alerts based on model predictions can then be forwarded to healthcare providers, establishing a proactive disease management loop.

5.2.11 Model Maintenance, Updating, and Drift Detection

And finally, continuity of a model with respect to time is very important. With the advancement of medical knowledge and the change in clinical profiles over time, periodic retraining and validation on updated datasets will be essential. Future work should create pipelines for concept/data drift detection and automated model weight update pipelines for when accuracy degrades.

This means that secure and privacy compliant data acquisition frameworks would have to be implemented along with version control for model artifacts. Federated learning could also be applied here to update the model across many institutions without actually moving sensitive patient data.

REFERENCES

[1] Jiang P., Li X., Shen H., Chen Y., Wang L., Chen H., Feng J., and Liu J., “A systematic review of deep learning-based cervical cytology screening: from cell identification to whole slide image analysis,” *Artificial Intelligence Review*, vol. 56, no. S2, p. 2687–2758, Oct. 2023. [Online]. Available: http://dx.doi.org/10.1007/s10462-023-10588-z

[2] Esteva A., Kuprel B., Novoa R. A., Ko J., Swetter S. M., Blau H. M., and Thrun S., “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, p. 115–118, Jan. 2017. [Online]. Available: http://dx.doi.org/10.1038/nature21056

[3] Bi W. L., Hosny A., Schabath M. B., Giger M. L., Birkbak N. J., Mehrtash A., Allison T., Arnaout O., Abbosh C., Dunn I. F., Mak R. H., Tamimi R. M., Tempany C. M., Swanton C., Hoffmann U., Schwartz L. H., Gillies R. J., Huang R. Y., and Aerts H. J. W. L., “Artificial intelligence in cancer imaging: Clinical challenges and applications,” *CA: A Cancer Journal for Clinicians*, vol. 69, no. 2, p. 127–157, Feb. 2019. [Online]. Available: http://dx.doi.org/10.3322/caac.21552

[4] Schmidt-Erfurth U., Sadeghipour A., Gerendas B. S., Waldstein S. M., and Bogunović H., “Artificial intelligence in retina,” *Progress in Retinal and Eye Research*, vol. 67, p. 1–29, Nov. 2018. [Online]. Available: http://dx.doi.org/10.1016/j.preteyeres.2018.07.004

[5] Akazawa M. and Hashimoto K., “Artificial intelligence in gynecologic cancers: Current status and future challenges – a systematic review,” *Artificial Intelligence in Medicine*, vol. 120, p. 102164, Oct. 2021. [Online]. Available: http://dx.doi.org/10.1016/j.artmed.2021.102164

[6] Salem N. M., Jack K. M., Gu H., Kumar A., Garcia M., Yang P., and Dinu V., “Machine and deep learning identified metabolites and clinical features associated with gallstone disease,” *Computer Methods and Programs*

in Biomedicine Update, vol. 3, p. 100106, 2023. [Online]. Available: http://dx.doi.org/10.1016/j.cmpbup.2023.100106

[7] Yao L., Mao C., and Luo Y., "Clinical text classification with rule-based features and knowledge-guided convolutional neural networks," in *2018 IEEE International Conference on Healthcare Informatics Workshop (ICHI-W)*. IEEE, Jun. 2018, p. 70–71. [Online]. Available: http://dx.doi.org/10.1109/ICHI-W.2018.00024

[8] Liew P.-L., Lee Y.-C., Lin Y.-C., Lee T.-S., Lee W.-J., Wang W., and Chien C.-W., "Comparison of artificial neural networks with logistic regression in prediction of gallbladder disease among obese patients," *Digestive and Liver Disease*, vol. 39, no. 4, p. 356–362, Apr. 2007. [Online]. Available: http://dx.doi.org/10.1016/j.dld.2007.01.003

[9] Obaid A. M., Turki A., Bellaaj H., Ksantini M., AlTaee A., and Alaerjan A., "Detection of gallbladder disease types using deep learning: An informative medical method," *Diagnostics*, vol. 13, no. 10, p. 1744, May 2023. [Online]. Available: http://dx.doi.org/10.3390/diagnostics13101744

[10] Obaid A. M., Turki A., Bellaaj H., and Ksantini M., "Diagnosis of gallbladder disease using artificial intelligence: A comparative study," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, Mar. 2024. [Online]. Available: http://dx.doi.org/10.1007/s44196-024-00431-w

[11] Thomas L., Manoj Kumar M. V., and Annappa, "A healthcare management using clinical decision support system," in *2018 3rd International Conference on Contemporary Computing and Informatics (IC3I)*. IEEE, Oct. 2018, p. 342–347. [Online]. Available: http://dx.doi.org/10.1109/IC3I44769.2018.9007272

[12] Blum J., Wood L., and Turner R., "Artificial intelligence in the detection of choledocholithiasis: a systematic review," *HPB*, vol. 27, no. 1, p. 1–9, Jan. 2025. [Online]. Available: http://dx.doi.org/10.1016/j.hpb.2024.09.009

[13] Vukicevic A. M., Stojadinovic M., Radovic M., Djordjevic M., Cirkovic B. A., Pejovic T., Jovicic G., and Filipovic N., "Automated development

of artificial neural networks for clinical purposes: Application for predicting the outcome of choledocholithiasis surgery," *Computers in Biology and Medicine*, vol. 75, p. 80–89, Aug. 2016. [Online]. Available: http://dx.doi.org/10.1016/j.compbiomed.2016.05.016

[14] Dalai C., Azizian J. M., Trieu H., Rajan A., Chen F. C., Dong T., Beaven S. W., and Tabibian J. H., "Machine learning models compared to existing criteria for noninvasive prediction of endoscopic retrograde cholangiopancreatography-confirmed choledocholithiasis," *Liver Research*, vol. 5, no. 4, p. 224–231, Dec. 2021. [Online]. Available: http://dx.doi.org/10.1016/j.livres.2021.10.001

[15] Floan Sachs G., Ourshalimian S., Jensen A. R., Kelley-Quon L. I., Padilla B. E., Shew S. B., Lofberg K. M., Smith C. A., Roach J. P., Pandya S. R., Russell K. W., and Ignacio R. C., "Machine learning to predict pediatric choledocholithiasis: A western pediatric surgery research consortium retrospective study," *Surgery*, vol. 174, no. 4, p. 934–939, Oct. 2023. [Online]. Available: http://dx.doi.org/10.1016/j.surg.2023.07.008

[16] Koçer S., Mohamed O., and Dündar O., "Disease detection in abdominal ct images using the yolov5 algorithm: A deep learning approach," in *2024 59th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST)*. IEEE, Jul. 2024, p. 1–4. [Online]. Available: http://dx.doi.org/10.1109/ICEST62335.2024.10639613

[17] Yu C.-J., Yeh H.-J., Chang C.-C., Tang J.-H., Kao W.-Y., Chen W.-C., Huang Y.-J., Li C.-H., Chang W.-H., Lin Y.-T., Sufriyana H., and Su E. C.-Y., "Lightweight deep neural networks for cholelithiasis and cholecystitis detection by point-of-care ultrasound," *Computer Methods and Programs in Biomedicine*, vol. 211, p. 106382, Nov. 2021. [Online]. Available: http://dx.doi.org/10.1016/j.cmpb.2021.106382

[18] Himi S. T., Monalisa N. T., Whaiduzzaman M., Barros A., and Uddin M. S., "Medai: A smartwatch-based application framework for the prediction of

common diseases using machine learning," *IEEE Access*, vol. 11, p. 12342–12359, 2023. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2023.3236002

[19] Zi Y., Cheng X., Mei T., Wang Q., Gao Z., and Yang H., "Research on intelligent system of medical image recognition and disease diagnosis based on big data," in *2024 IEEE 2nd International Conference on Image Processing and Computer Applications (ICIPCA)*. IEEE, Jun. 2024, p. 825–830. [Online]. Available: http://dx.doi.org/10.1109/ICIPCA61593.2024.10709204

[20] Yao C., Wu S., Liu Z., and Li P., "A deep learning model for predicting chemical composition of gallstones with big data in medical internet of things," *Future Generation Computer Systems*, vol. 94, p. 140–147, May 2019. [Online]. Available: http://dx.doi.org/10.1016/j.future.2018.11.011

[21] Zhou S., Ye L., Huang Y., Valsecchi C., Liu Y., Shao L., Liu J., He T., Liu L., and Fan M., "Rapid diagnosis and recurrence prediction of choledocholithiasis disease using raw bile with machine learning assisted sers," *Talanta*, vol. 282, p. 126979, Jan. 2025. [Online]. Available: http://dx.doi.org/10.1016/j.talanta.2024.126979

[22] Turki A., Obaid A. M., Bellaaj H., Ksantini M., and AlTaee A., "Uidatagb: Multi-class ultrasound images dataset for gallbladder disease detection," *Data in Brief*, vol. 54, p. 110426, Jun. 2024. [Online]. Available: http://dx.doi.org/10.1016/j.dib.2024.110426

[23] Vanitha M., Ganesh Babu T. R., Praveena R., and Sathyapriya M., "Segmentation and classification of gallstone in ultrasound images of gall bladder using support vector machine," in *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, Oct. 2022, p. 1–6. [Online]. Available: http://dx.doi.org/10.1109/ICCCNT54827.2022.9984361

[24] Yanto M., Yunus Y., and Tajuddin M., "Unsupervised multi-stage segmentation in image identification of gallstone disease," in *2023 International Conference on Information Technology Research and Innovation*

(ICITRI). IEEE, Aug. 2023, p. 38–42. [Online]. Available: http://dx.doi.org/10.1109/ICITRI59340.2023.10250105

[25] Song T., Meng F., Rodriguez-Paton A., Li P., Zheng P., and Wang X., "U-next: A novel convolution neural network with an aggregation u-net architecture for gallstone segmentation in ct images," *IEEE Access*, vol. 7, p. 166823–166832, 2019. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2019.2953934

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

[27] Kira K. and Rendell L. A., "A practical approach to feature selection," in *Machine learning proceedings 1992*. Elsevier, 1992, pp. 249–256.

[28] Dokeroglu T., Deniz A., and Kiziloz H. E., "A comprehensive survey on recent metaheuristics for feature selection," *Neurocomputing*, vol. 494, p. 269–296, Jul. 2022. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2022.04.083

[29] Peng H., Long F., and Ding C., "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, p. 1226–1238, Aug. 2005. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2005.159

[30] Agar M., Aydin S., Cakmak M., Koc M., and Togacar M., "Detection of thymoma disease using mrmr feature selection and transformer models," *Diagnostics*, vol. 14, no. 19, p. 2169, Sep. 2024. [Online]. Available: http://dx.doi.org/10.3390/diagnostics14192169

[31] Toğaçar M., Ergen B., and Cömert Z., "Detection of lung cancer on chest ct images using minimum redundancy maximum relevance feature selection method with convolutional neural networks," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 1, p. 23–39, Jan. 2020. [Online]. Available: http://dx.doi.org/10.1016/j.bbe.2019.11.004

[32] Buś S., Jędrzejewski K., and Guzik P., "Using minimum redundancy maximum relevance algorithm to select minimal sets of heart rate variability parameters for atrial fibrillation detection," *Journal of Clinical Medicine*, vol. 11, no. 14, p. 4004, Jul. 2022. [Online]. Available: http://dx.doi.org/10.3390/jcm11144004

[33] Pearson K., "X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 50, no. 302, p. 157–175, Jul. 1900. [Online]. Available: http://dx.doi.org/10.1080/14786440009463897

[34] Korial A. E., Gorial I. I., and Humaidi A. J., "An improved ensemble-based cardiovascular disease detection system with chi-square feature selection," *Computers*, vol. 13, no. 6, p. 126, May 2024. [Online]. Available: http://dx.doi.org/10.3390/computers13060126

[35] Çalışkan A., "Diagnosis of malaria disease by integrating chi-square feature selection algorithm with convolutional neural networks and autoencoder network," *Transactions of the Institute of Measurement and Control*, vol. 45, no. 5, pp. 975–985, 2023.

[36] *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 10, no. 3, p. 2050–2060, Jun. 2021. [Online]. Available: http://dx.doi.org/10.30534/ijatcse/2021/801032021

[37] Fisher S. R. A., "Statistical methods for research workers," p. xxxv–xxxviii, Apr. 1990. [Online]. Available: http://dx.doi.org/10.1093/oso/9780198522294.002.0003

[38] Vrigazova B. P., "Detection of malignant and benign breast cancer using the anova-bootstrap-svm," *Journal of Data and Information Science*, vol. 5, no. 2, p. 62–75, Apr. 2020. [Online]. Available: http://dx.doi.org/10.2478/jdis-2020-0012

[39] Chowdhury M. N. R., Zhang W., and Akilan T., “Anova-based automatic attribute selection and a predictive model for heart disease prognosis,” 2022. [Online]. Available: https://arxiv.org/abs/2208.00296

[40] Nasiri H. and Alavi S. A., “A novel framework based on deep learning and anova feature selection method for diagnosis of covid-19 cases from chest x-ray images,” *Computational Intelligence and Neuroscience*, vol. 2022, p. 1–11, Jan. 2022. [Online]. Available: http://dx.doi.org/10.1155/2022/4694567

[41] Chan Y. and Walmsley R. P., “Learning and understanding the kruskal-wallis one-way analysis-of-variance-by-ranks test for differences among three or more independent groups,” *Physical Therapy*, vol. 77, no. 12, p. 1755–1761, Dec. 1997. [Online]. Available: http://dx.doi.org/10.1093/ptj/77.12.1755

[42] Jamil M. A. and Khanam S., “Influence of one-way anova and kruskal–wallis based feature ranking on the performance of ml classifiers for bearing fault diagnosis,” *Journal of Vibration Engineering & Technologies*, vol. 12, no. 3, p. 3101–3132, Jun. 2023. [Online]. Available: http://dx.doi.org/10.1007/s42417-023-01036-x

[43] Acar E. F. and Sun L., “A generalized kruskal–wallis test incorporating group uncertainty with application to genetic association studies,” *Biometrics*, vol. 69, no. 2, p. 427–435, Feb. 2013. [Online]. Available: http://dx.doi.org/10.1111/biom.12006

[44] Sherwani R. A. K., Shakeel H., Awan W. B., Faheem M., and Aslam M., “Analysis of covid-19 data using neutrosophic kruskal wallis h test,” *BMC Medical Research Methodology*, vol. 21, no. 1, Oct. 2021. [Online]. Available: http://dx.doi.org/10.1186/s12874-021-01410-x

[45] Mirjalili S., Mirjalili S. M., and Lewis A., “Grey wolf optimizer,” *Advances in Engineering Software*, vol. 69, p. 46–61, Mar. 2014. [Online]. Available: http://dx.doi.org/10.1016/j.advengsoft.2013.12.007

[46] Niu S., Zhou Y., Li Z., Huang S., and Zhou Y., "An improved grey wolf optimization algorithm for heart disease prediction," 2024. [Online]. Available: https://arxiv.org/abs/2401.11669

[47] Er B. and Fidan U., "Koroner arter hastalığı sınıflandırılmasında destek vektör makinelerinin gri kurt optimizasyonuna dayalı Özellik seçim yöntemi ile geliştirilmesi," *Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi*, vol. 5, no. 1, p. 37–44, Jun. 2024. [Online]. Available: http://dx.doi.org/10.53608/estudambilisim.1409734

[48] Ochoa-Ornelas R., Gudiño-Ochoa A., and García-Rodríguez J. A., "A hybrid deep learning and machine learning approach with mobile-efficientnet and grey wolf optimizer for lung and colon cancer histopathology classification," *Cancers*, vol. 16, no. 22, p. 3791, Nov. 2024. [Online]. Available: http://dx.doi.org/10.3390/cancers16223791

[49] Heidari A. A., Mirjalili S., Faris H., Aljarah I., Mafarja M., and Chen H., "Harris hawks optimization: Algorithm and applications," *Future Generation Computer Systems*, vol. 97, p. 849–872, Aug. 2019. [Online]. Available: http://dx.doi.org/10.1016/j.future.2019.02.028

[50] Aly S. A., Bakhiet A., and Balat M., "Diagnosis of malignant lymphoma cancer using hybrid optimized techniques based on dense neural networks," 2024. [Online]. Available: https://arxiv.org/abs/2410.06974

[51] Wang M., Chen L., Heidari A. A., and Chen H., "Fireworks explosion boosted harris hawks optimization for numerical optimization: Case of classifying the severity of covid-19," *Frontiers in Neuroinformatics*, vol. 16, Jan. 2023. [Online]. Available: http://dx.doi.org/10.3389/fninf.2022.1055241

[52] Dokeroglu T., "A new parallel multi-objective harris hawk algorithm for predicting the mortality of covid-19 patients," *PeerJ Computer Science*, vol. 9, p. e1430, Jun. 2023. [Online]. Available: http://dx.doi.org/10.7717/peerj-cs.1430

[53] Kennedy J. and Eberhart R., "Particle swarm optimization," in *Proceedings of ICNN'95 - International Conference on Neural Networks*, ser.

ICNN-95, vol. 4. IEEE, p. 1942–1948. [Online]. Available: http://dx.doi.org/10.1109/ICNN.1995.488968

[54] Yang G., Li W., Xie W., Wang L., and Yu K., "An improved binary particle swarm optimization algorithm for clinical cancer biomarker identification in microarray data," *Computer Methods and Programs in Biomedicine*, vol. 244, p. 107987, Feb. 2024. [Online]. Available: http://dx.doi.org/10.1016/j.cmpb.2023.107987

[55] Pramanik R., Sarkar S., and Sarkar R., "An adaptive and altruistic pso-based deep feature selection method for pneumonia detection from chest x-rays," 2022. [Online]. Available: https://arxiv.org/abs/2208.03558

[56] KILIÇARSLAN S., ADEM K., and CÖMERT O., "Parçacık sürü optimizasyonu kullanılarak boyutu azaltılmış mikrodizi verileri Üzerinde makine Öğrenmesi yöntemleri ile prostat kanseri teşhisi," *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, vol. 7, no. 1, p. 769–777, Jan. 2019. [Online]. Available: http://dx.doi.org/10.29130/dubited.464092

[57] Mirjalili S. and Lewis A., "The whale optimization algorithm," *Advances in Engineering Software*, vol. 95, p. 51–67, May 2016. [Online]. Available: http://dx.doi.org/10.1016/j.advengsoft.2016.01.008

[58] Atimbire S. A., Appati J. K., and Owusu E., "Empirical exploration of whale optimisation algorithm for heart disease prediction," *Scientific Reports*, vol. 14, no. 1, Feb. 2024. [Online]. Available: http://dx.doi.org/10.1038/s41598-024-54990-1

[59] Nadimi-Shahraki M. H., Zamani H., and Mirjalili S., "Enhanced whale optimization algorithm for medical feature selection: A covid-19 case study," *Computers in Biology and Medicine*, vol. 148, p. 105858, Sep. 2022. [Online]. Available: http://dx.doi.org/10.1016/j.compbiomed.2022.105858

[60] Majid A., Alrasheedi M. A., Alharbi A. A., Allohibi J., and Lee S.-W., "Modified whale optimization algorithm for multiclass skin cancer classification," *Mathematics*, vol. 13, no. 6, p. 929, Mar. 2025. [Online]. Available: http://dx.doi.org/10.3390/math13060929

[61] Clarke M. R. B., Duda R. O., and Hart P. E., "Pattern classification and scene analysis." *Journal of the Royal Statistical Society. Series A (General)*, vol. 137, no. 3, p. 442, 1974. [Online]. Available: http://dx.doi.org/10.2307/2344977

[62] Asan M. E., Taşkın H., Alemdar M., and Capoglu R., "Tiroit kanseri hastalık tanısında lojistik regresyon kullanımı," *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, vol. 39, no. 3, p. 1509–1524, May 2024. [Online]. Available: http://dx.doi.org/10.17341/gazimmfd.1253193

[63] ALTUN S. and Alkan A., "Mr spektroskopi temelli beyin tümörü teşhisinde lojistik regresyon uygulaması," *Kahramanmaraş Sütçü İmam Üniversitesi Mühendislik Bilimleri Dergisi*, vol. 22, no. 1, p. 10–18, Apr. 2019. [Online]. Available: http://dx.doi.org/10.17780/ksujes.483214

[64] ŞAMKAR H., YILDIRIM A. G., and DELİBAŞ O., "Determining the risk factors causing cancer with logistic regression analysis," *Alphanumeric Journal*, vol. 4, no. 2, p. 205–205, Dec. 2016. [Online]. Available: http://dx.doi.org/10.17093/alphanumeric.277744

[65] Minsky M., "Steps toward artificial intelligence," *Proceedings of the IRE*, vol. 49, no. 1, p. 8–30, Jan. 1961. [Online]. Available: http://dx.doi.org/10.1109/JRPROC.1961.287775

[66] Karabatak M., "A new classifier for breast cancer detection based on naïve bayesian," *Measurement*, vol. 72, p. 32–36, Aug. 2015. [Online]. Available: http://dx.doi.org/10.1016/j.measurement.2015.04.028

[67] Bhagya Shree S. R. and Sheshadri H. S., "Diagnosis of alzheimer's disease using naive bayesian classifier," *Neural Computing and Applications*, vol. 29, no. 1, p. 123–132, Jul. 2016. [Online]. Available: http://dx.doi.org/10.1007/s00521-016-2416-3

[68] Do B. H., Langlotz C., and Beaulieu C. F., "Bone tumor diagnosis using a naïve bayesian model of demographic and radiographic features," *Journal of Digital Imaging*, vol. 30, no. 5, p. 640–647, Jul. 2017. [Online]. Available: http://dx.doi.org/10.1007/s10278-017-0001-7

[69] Boser B. E., Guyon I. M., and Vapnik V. N., "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, ser. COLT92. ACM, Jul. 1992, p. 144–152. [Online]. Available: http://dx.doi.org/10.1145/130385.130401

[70] Sandhy Y. a., "Prediction of heart diseases using support vector machine," *International Journal for Research in Applied Science and Engineering Technology*, vol. 8, no. 2, p. 126–135, Feb. 2020. [Online]. Available: http://dx.doi.org/10.22214/ijraset.2020.2021

[71] Karapinar Senturk Z., "Early diagnosis of parkinson's disease using machine learning algorithms," *Medical Hypotheses*, vol. 138, p. 109603, May 2020. [Online]. Available: http://dx.doi.org/10.1016/j.mehy.2020.109603

[72] Wang H., Zheng B., Yoon S. W., and Ko H. S., "A support vector machine-based ensemble algorithm for breast cancer diagnosis," *European Journal of Operational Research*, vol. 267, no. 2, p. 687–699, Jun. 2018. [Online]. Available: http://dx.doi.org/10.1016/j.ejor.2017.12.001

[73] Quinlan J. R., "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, p. 81–106, Mar. 1986. [Online]. Available: http://dx.doi.org/10.1007/BF00116251

[74] Maji S. and Arora S., *Decision Tree Algorithms for Prediction of Heart Disease*. Springer Singapore, Aug. 2018, p. 447–454. [Online]. Available: http://dx.doi.org/10.1007/978-981-13-0586-3_45

[75] Nahar N. and Ara F., "Liver disease prediction by using different decision tree techniques," *International Journal of Data Mining & Knowledge Management Process*, vol. 8, no. 2, p. 01–09, Mar. 2018. [Online]. Available: http://dx.doi.org/10.5121/ijdkp.2018.8201

[76] T.R. M., V. V. K., Bhardwaj R., Khan S. B., Alkhaldi N. A., Victor N., and Verma A., "An artificial intelligence-based decision support system for early and accurate diagnosis of parkinson's disease," *Decision Analytics Journal*, vol. 10, p. 100381, Mar. 2024. [Online]. Available: http://dx.doi.org/10.1016/j.dajour.2023.100381

[77] Cover T. and Hart P., "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, p. 21–27, Jan. 1967. [Online]. Available: http://dx.doi.org/10.1109/TIT.1967.1053964

[78] Rahmat D., Putra A. A., Hamrin, and Setiawan A. W., "Heart disease prediction using k-nearest neighbor," in *2021 International Conference on Electrical Engineering and Informatics (ICEEI)*. IEEE, Oct. 2021, p. 1–6. [Online]. Available: http://dx.doi.org/10.1109/ICEEI52609.2021.9611110

[79] Chen H.-L., Huang C.-C., Yu X.-G., Xu X., Sun X., Wang G., and Wang S.-J., "An efficient diagnosis system for detection of parkinson's disease using fuzzy k-nearest neighbor approach," *Expert Systems with Applications*, vol. 40, no. 1, p. 263–271, Jan. 2013. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2012.07.014

[80] Ali A., Alrubei M. A. T., Hassan L. F. M., Al-Ja'afari M. A. M., and Abdulwahed S. H., "Diabetes diagnosis based on knn," *IIUM Engineering Journal*, vol. 21, no. 1, p. 175–181, Jan. 2020. [Online]. Available: http://dx.doi.org/10.31436/iiumej.v21i1.1206

[81] Friedman J. H., "Greedy function approximation: A gradient boosting machine." *The Annals of Statistics*, vol. 29, no. 5, Oct. 2001. [Online]. Available: http://dx.doi.org/10.1214/aos/1013203451

[82] Kumar K. L. and Reddy B. E., "Heart disease detection system using gradient boosting technique," in *2021 International Conference on Computing Sciences (ICCS)*. IEEE, Dec. 2021, p. 228–233. [Online]. Available: http://dx.doi.org/10.1109/ICCS54944.2021.00052

[83] Karabayir I., Goldman S. M., Pappu S., and Akbilgic O., "Gradient boosting for parkinson's disease diagnosis from voice recordings," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, Sep. 2020. [Online]. Available: http://dx.doi.org/10.1186/s12911-020-01250-7

[84] Ahmed H., Soliman H., and Elmogy M., "Early detection of alzheimer's disease using single nucleotide polymorphisms analysis based on gradient boosting tree,"

Computers in Biology and Medicine, vol. 146, p. 105622, Jul. 2022. [Online]. Available: http://dx.doi.org/10.1016/j.compbiomed.2022.105622

[85] Freund Y. and Schapire R. E., "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, p. 119–139, Aug. 1997. [Online]. Available: http://dx.doi.org/10.1006/jcss.1997.1504

[86] Asra T., Setiadi A., Safudin M., Lestari E. W., Hardi N., and Alamsyah D. P., "Implementation of adaboost algorithm in prediction of chronic kidney disease," in *2021 7th International Conference on Engineering, Applied Sciences and Technology (ICEAST)*. IEEE, Apr. 2021, p. 264–268. [Online]. Available: http://dx.doi.org/10.1109/ICEAST52143.2021.9426291

[87] Mahesh T. R., Dhilip Kumar V., Vinoth Kumar V., Asghar J., Geman O., Arulkumaran G., and Arun N., "Adaboost ensemble methods using k-fold cross validation for survivability with the early detection of heart disease," *Computational Intelligence and Neuroscience*, vol. 2022, p. 1–11, Apr. 2022. [Online]. Available: http://dx.doi.org/10.1155/2022/9005278

[88] Ganachari S. and Battula S. R., *Stroke Disease Prediction Using Adaboost Ensemble Learning Technique*. Springer Nature Singapore, 2023, p. 247–260. [Online]. Available: http://dx.doi.org/10.1007/978-981-99-2100-3_21

[89] Ho T. K., "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, ser. ICDAR-95, vol. 1. IEEE Comput. Soc. Press, p. 278–282. [Online]. Available: http://dx.doi.org/10.1109/ICDAR.1995.598994

[90] Singh Y. K., Sinha N., and Singh S. K., *Heart Disease Prediction System Using Random Forest*. Springer Singapore, 2017, p. 613–623. [Online]. Available: http://dx.doi.org/10.1007/978-981-10-5427-3_63

[91] Khalilia M., Chakraborty S., and Popescu M., "Predicting disease risks from highly imbalanced data using random forest," *BMC Medical Informatics*

and Decision Making, vol. 11, no. 1, Jul. 2011. [Online]. Available: http://dx.doi.org/10.1186/1472-6947-11-51

[92] Islam M. T., Raihan M., Farzana F., Aktar N., Ghosh P., and Kabiraj S., "Typical and non-typical diabetes disease prediction using random forest algorithm," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, Jul. 2020, p. 1–6. [Online]. Available: http://dx.doi.org/10.1109/ICCCNT49239.2020.9225430

[93] Rosenblatt F., "The perceptron: A probabilistic model for information storage and organization in the brain." *Psychological Review*, vol. 65, no. 6, p. 386–408, 1958. [Online]. Available: http://dx.doi.org/10.1037/h0042519

[94] Sonawane J. S. and Patil D. R., "Prediction of heart disease using multilayer perceptron neural network," in *International Conference on Information Communication and Embedded Systems (ICICES2014)*. IEEE, Feb. 2014, p. 1–6. [Online]. Available: http://dx.doi.org/10.1109/ICICES.2014.7033860

[95] Selwal A. and Raoof I., "A multi-layer perceptron based improved thyroid disease prediction system," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 17, no. 1, p. 524, Jan. 2020. [Online]. Available: http://dx.doi.org/10.11591/ijeecs.v17.i1.pp524-532

[96] Yildirim P., "Chronic kidney disease prediction on imbalanced data by multilayer perceptron: Chronic kidney disease prediction," in *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*. IEEE, Jul. 2017, p. 193–198. [Online]. Available: http://dx.doi.org/10.1109/COMPSAC.2017.84

APPENDICES

Appendix A: Features of the Dataset.

Appendix B: List of library and module versions used.

Appendix A – Features of the Dataset

Table A.1 Features of the Dataset

Features	Explanation
Age	Age of the person
Gender	Gender of the person
Comorbidity	Concomitant diseases
Coronary Artery Disease (CAD)	Cardiovascular disease
Hypothyroidism	Underactive thyroid gland
Hyperlipidemia	High levels of fat in the blood
Diabetes Mellitus (DM)	High blood sugar
Height	Height is the length
Weight	Body weight
Body Mass Index (BMI)	Weight for height ratio
Total Body Water (TBW)	Total water in the body
Extracellular Water (ECW)	It is extracellular water
Intracellular Water (ICW)	It is intracellular water
Extracellular Fluid / Total Body Water (ECF/TBW)	Extracellular water content
Total Body Fat Ratio (TBFR)	Total fat content
Lean Mass (LM)	Lean body mass
Protein	Essential building blocks for the body
Visceral Fat Rating (VFR)	Visceral organ fat level
Bone Mass (BM)	Bone weight
Muscle Mass (MM)	Amount of muscle
Obesity	Excessive adiposity
Total Fat Content (TFC)	Total fat amount
Visceral Fat Area (VFA)	Inner adipose tissue area
Visceral Muscle Area (VMA)	Inner muscle area
Hepatic Fat Accumulation (HFA)	Accumulation of fat in the liver
Glucose	Blood sugar
Total Cholesterol (TC)	It is total cholesterol
Low Density Lipoprotein (LDL)	Bad cholesterol
High Density Lipoprotein (HDL)	Good cholesterol
Triglyceride	Type of fat found in the blood
Aspartat Aminotransferaz (AST)	Type of liver enzyme
Alanin Aminotransferaz (ALT)	An enzyme related to the liver
Alkaline Phosphatase (ALP)	Type of liver and bone enzyme
Creatinine	Kidney function indicator
Glomerular Filtration Rate (GFR)	Kidney filtration rate
C-reactive Protein (CRP)	Inflammation indicator
Hemoglobin	Protein in the blood that carries oxygen
Vitamin D	Essential vitamin for bone health

Appendix B – List of library and module versions used

Table B.1 List of library and module versions used

Library / Module	Version
MATLAB	R2023b
Python	3.11.5
Jupyter Notebook	7.3.3
pandas	2.1.1
numpy	1.26.0
scikit-learn	1.6.1
openpyxl	3.1.2