A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF ÇANKIRI KARATEKIN UNIVERSITY

ANALYSIS OF CANCER DATASET WITH STATISTICAL LEARNING

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRONICS AND COMPUTER ENGINEERING

BY

ASMAA SALIM HUSSAIEN ALWAZY

ÇANKIRI

2024

ANALYSIS OF CANCER DATASET WITH STATISTICAL LEARNING

By Asmaa Salim Hussaien ALWAZY

January 2024

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science

Advisor : Assoc. Prof. Dr. Selim BUYRUKOĞLU
Co-Advisor : Asst. Prof. Dr. Gonca BUYRUKOĞLU

Examining Committee Members:

Chairman : Assoc. Prof. Dr. Selim BUYRUKOĞLU
Electronics and Computer Engineering
Çankırı Karatekin University

Member : Assoc. Prof. Dr. Serkan SAVAŞ
Computer Engineering
Kırıkkale University

Member : Asst. Prof. Dr. Taha ETEM
Electronics and Computer Engineering
Çankırı Karatekin University

Approved for the Graduate School of Natural and Applied Sciences

Prof. Dr. Hamit ALYAR
Director of Graduate School

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Asmaa Salim Hussaien ALWAZY

ABSTRACT

ANALYSIS OF CANCER DATASET WITH STATISTICAL LEARNING

Asmaa Salim Hussaien ALWAZY

Master of Science in Electronics and Computer Engineering

Advisor: Assoc. Prof. Dr. Selim BUYRUKOĞLU

Co-Advisor: Asst. Prof. Dr. Gonca BUYRUKOĞLU

January 2024

Cancer continues to pose a significant global health challenge, underscoring the criticality of early and accurate diagnosis for enhancing treatment outcomes and patient well-being. The classification of cancer types assumes a pivotal role in tailoring treatment plans, minimizing unnecessary procedures, and optimizing therapeutic success. This thesis presents an extensive analysis of statistical learning algorithms and machine learning (ML) algorithms on breast cancer, lung cancer, and prostate cancer datasets. The primary objective was to evaluate the algorithms' performance in distinguishing between benign and malignant samples across diverse cancer types. To ensure robust and reliable results, a comprehensive steps of preprocessing techniques was implemented, encompassing data cleaning to address null values and duplicate records, data scaling for feature normalization, random over-sampling to tackle class imbalance, and an 80:20 data splitting ratio for training and testing. Additionally, cross-validation was employed to assess model generalization and robustness. The paramount importance of accurately diagnosing cancer types lies in its potential to significantly impact patient outcomes and guide treatment strategies. The results showcased impressive accuracies ranging from 95.8% using ridge logistic regression to 97.2% using lasso logistic regression for breast cancer. Similarly, ML algorithms, such as Decision Tree, SVM, Random Forest, and XGBoost, achieved accuracies between 93% using random forest to 98.6% using XGBoost for breast cancer. Additionally, lung cancer statistical learning algorithms demonstrated accuracies between 93.75% using Ridge regression to 96.87% using Lasso regression, while ML algorithms achieved accuracies from 95.83% using Decision tree

to 98.95% using Random forest. For prostate cancer, statistical learning algorithms achieved accuracies between 74.11% using ElasticNet regression to 77.64% using Lasso regression, and ML algorithms achieved accuracies ranging from 63.53% using Decision tree to 75.29% using SVM. These findings underscore the effectiveness of both statistical learning and ML algorithms in cancer classification, affirming their potential applicability in real-world scenarios to advance cancer detection and diagnosis.

2024, 66 pages

Keywords: Cancer diagnosis, Statistical learning algorithms, Machine learning algorithms

ÖZET

KANSER VERİ SETİNİN İSTATİSTİKSEL ÖĞRENME İLE ANALİZİ

Asmaa Salim Hussaien ALWAZY
Elektronik ve Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Doç. Dr. Selim BUYRUKOĞLU
Eş Danışman: Dr. Öğr. Üyesi Gonca BUYRUKOĞLU
Ocak 2024

Kanser, önemli bir küresel sağlık sorunu olarak kalmaya devam ediyor, bu da erken ve doğru teşhisin tedavi sonuçlarını ve hasta iyiliğini artırmadaki kritik önemini vurguluyor. Kanser türlerinin sınıflandırılması, tedavi planlarını kişiselleştirmede, gereksiz işlemleri en aza indirmede ve terapötik başarıyı optimize etmede merkezi bir rol üstleniyor. Bu tez, meme kanseri, akciğer kanseri ve prostat kanseri veri setleri üzerinde istatistiksel öğrenme algoritmaları ve makine öğrenmesi (ML) algoritmalarının kapsamlı bir analizini sunuyor. Ana hedef, çeşitli kanser türleri arasında iyi huylu ve kötü huylu örnekleri ayırt etmede algoritmaların performansını değerlendirmekti. Sağlam ve güvenilir sonuçlar sağlamak için, boş değerlerle ve çift kayıtlarla ilgilenmek üzere veri temizleme, özellik normalizasyonu için veri ölçeklendirme, sınıf dengesizliğiyle başa çıkmak için rastgele fazla örneklem yapma ve eğitim ve test için %80:%20 veri bölme oranını kapsayan kapsamlı bir ön işleme teknikleri adımı uygulandı. Ayrıca, modelin genelleştirme ve sağlamlığını değerlendirmek için çapraz doğrulama kullanıldı. Kanser türlerini doğru bir şekilde teşhis etmenin son derece önemli olması, hasta sonuçları üzerinde önemli bir etki yapma potansiyeline ve tedavi stratejilerini yönlendirme kapasitesine sahip olmasından kaynaklanıyor. Sonuçlar, meme kanseri için sırtlı lojistik regresyon kullanılarak %95.8'den lasso lojistik regresyon kullanılarak %97.2'ye kadar olan etkileyici doğruluk oranlarını sergiledi. Benzer şekilde, Karar Ağacı, SVM, Rastgele Orman ve XGBoost gibi ML algoritmaları, meme kanseri için Rastgele Orman kullanılarak %93'ten XGBoost kullanılarak %98.6'ya kadar doğruluk oranları elde etti. Ayrıca, akciğer kanseri için istatistiksel öğrenme algoritmaları, Sırtlı regresyon kullanılarak %93.75'ten Lasso regresyon kullanılarak %96.87'ye kadar doğruluk oranlarını gösterirken, ML

algoritmaları Karar Ağacı kullanılarak %95.83'ten Rastgele Orman kullanılarak %98.95'e kadar doğruluk oranları elde etti. Prostat kanseri için, istatistiksel öğrenme algoritmaları ElastikNet regresyonu kullanılarak %74.11'den Lasso regresyonu kullanılarak %77.64'e kadar, ve ML algoritmaları Karar Ağacı kullanılarak %63.53'ten SVM kullanılarak %75.29'a kadar doğruluk oranları elde etti. Bu bulgular, kanser sınıflandırmasında hem istatistiksel öğrenme hem de ML algoritmalarının etkinliğini vurguluyor ve kanser tespiti ve teşhisinde gerçek dünya senaryolarında uygulanabilirlik potansiyellerini doğruluyor.

2024, 66 sayfa

Anahtar Kelimeler: Kanser tanısı, İstatistiksel öğrenme algoritmaları, Makine öğrenme algoritmaları

PREFACE AND ACKNOWLEDGEMENTS

I would like to thank my thesis advisor, Assoc. Prof. Dr. Selim BUYRUKOĞLU, for his patience, guidance and understanding, and I would like to express my heartfelt gratitude to Asst. Prof. Dr. Gonca BUYRUKOĞLU, for her expertise and encouragement.

Asmaa Salim Hussaien ALWAZY
Çankırı-2024

CONTENTS

LIST OF ABBREVIATIONS

AI	Artificial intelligence
ANN	Artificial neural network
BCD	Breast cancer dataset
CNS	Central nervous system
CT	Computed tomography scan
CPH	Cox proportional hazards
CV	Cross validation
DFS	Disease free survival
XGB	Extreme gradient boosting
ELM	Extreme learning machine
FN	False negative
FP	False positive
HCC	Hepatocellular carcinoma
ID	Identifier
KNN	K-nearest neighbors
LR	Logistic regression
ML	Machine learning
NB	Naive bayes
OS	Overall survival
PCA	Principal component analysis
RBF	Radial basis function
RF	Random forest
RSFs	Random survival forests
SVC	Support vector classifier
SVM	Support vector machine
SMOTE	Synthetic minority oversampling technique
TN	True negative
TP	True positive
TAS	Tumor aggression score
TNM	Tymor, nodes, metastasis
WHO	Worls health organization

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Studying cancer holds significant importance due to its status as one of the major causes of global fatalities. Cancer is defined as a condition marked by the uncontrolled proliferation and dissemination of cells within the human body, with the potential to initiate in virtually any of the trillions of cells comprising the human anatomy (Ferlay *et al.* 2020).

And according to another definition, cancer is a collection of over a hundred distinct disorders. It can form virtually anywhere on the body.

Cancer is a hereditary condition resulting from genetic or epigenetic abnormalities in somatic cells, marked by abnormal cell proliferation that can extend to other organs. In 2018, the global count of cancer cases reached 18 million, with 9.5 million instances in men, 8.5 million in women, and 9.6 million fatalities. Prostate, breast, lung, stomach, colorectal, and non-melanoma skin cancers rank among the most widespread types globally, with over 100 varieties affecting individuals. The impact of cancer continues to escalate daily. Tobacco contributes to 22% of cases, diseases like HIV, hepatitis B, and Epstein-Barr virus account for 15%, and factors such as poor diet, obesity, excessive alcohol consumption, and exposure to ionizing radiation contribute to 10%. Various cancer-causing factors are under scrutiny. Cancer types include Carcinomas, originating in tissues or skin covering glands and internal organs, resulting in solid tumors, such as breast cancer, prostate cancer, colorectal cancer, and lung cancer (National Cancer Institute 2022).

Sarcomas initiate in the body's connective and supportive tissues, with the potential to develop in nerves, tendons, joints, fat, blood vessels, bone, lymphatic vessels, muscles, and cartilage. Patients are typically treated with chemotherapy, radiation therapy, immunotherapy, surgery, hormone therapy, and combinations thereof. Stem cell transplant is likewise the most effective treatment for cancer.

Today, a significant amount of research is being conducted on precision medicine to improve the future techniques (methods) detection and early detection of cancer for increase the chance of survival. After reading so many papers that used statical approaches to detect cancer. There are many datasets that are used to detect cancer (National Cancer Institute 2022).

1.1 Motivation

- Early detection: Early detection of cancer is important for successful treatment and increased survival rates. Machine Learning (ML) and statistical learning algorithms can be used to identify patterns in data that may indicate the presence of cancer, allowing for earlier detection and treatment.
- Accuracy: ML and statistical learning algorithms possess the capability to enhance the precision of cancer detection. By analyzing large amounts of data, these algorithms can identify subtle differences between cancerous and non-cancerous cells that may not be detectable by human analysis.
- Personalized treatment: Different types of cancer require different treatments, and individual patients may respond differently to the same treatment. Machine learning and statistical learning algorithms can help to identify which treatments are most effective for different types of cancer and for different individuals, improving treatment outcomes.
- Efficiency: ML and statistical learning algorithms can help to streamline the diagnosis and treatment process, reducing the time and resources needed to diagnose and treat cancer.
- Future advancements: As more data becomes available, machine learning and statistical learning algorithms will become even more powerful tools for cancer diagnosis and treatment. Ongoing research in this area has the potential to significantly improve cancer outcomes in the future.

1.2 Aim and Objectives

Our aim is to develop and validate ML and statistical learning algorithms for accurate cancer diagnosing and diagnosis based on molecular and genomic features. We can also conclude the objective in the following points:

- To compile and preprocess large and diverse datasets from various cancer types.
- To develop and optimize ML like support vector machines (SVM), random forests (RF), and Neural Networks (NN), for the classification of different cancer types based on molecular and genomic features.
- To investigate the performance of ML algorithms for cancer diagnosis and classification on independent datasets.
- To compare the accuracy of ML algorithms for cancer classification and diagnosis with existing methods.
- To develop and validate statistical learning algorithms, such as regression models to predict patient outcomes and responses to different cancer treatments based on molecular and genomic features.

1.3 Research Questions

- Research Question 1 : How can ML algorithms help to accurately classify different types of cancer cells ?
- Research Question 2: Can statistical learning algorithms be used to predict the effectiveness of different cancer treatments based on patient-specific characteristics, and how accurate are these predictions?
- Research Question 3 : How can we measure the robustness of the model ?
- Research Question 4 : What are the most effective approaches for cancer diagnosis and classification and what impact does this have on treatment outcomes?

1.4 Research Organization

In the rest of this thesis, chapter 2 reviews the relevant literature on cancer classification and diagnosis. Chapter 3 outlines the methodology used for cancer classification and diagnosis, including data preprocessing, feature selection, and algorithm selection. Chapter 4 represents the results of the proposed work and discussion. Finally, Chapter 5 summarizes the main findings and future research directions of the study and finally the references. Figure 1.1 represents the overall theis organization.

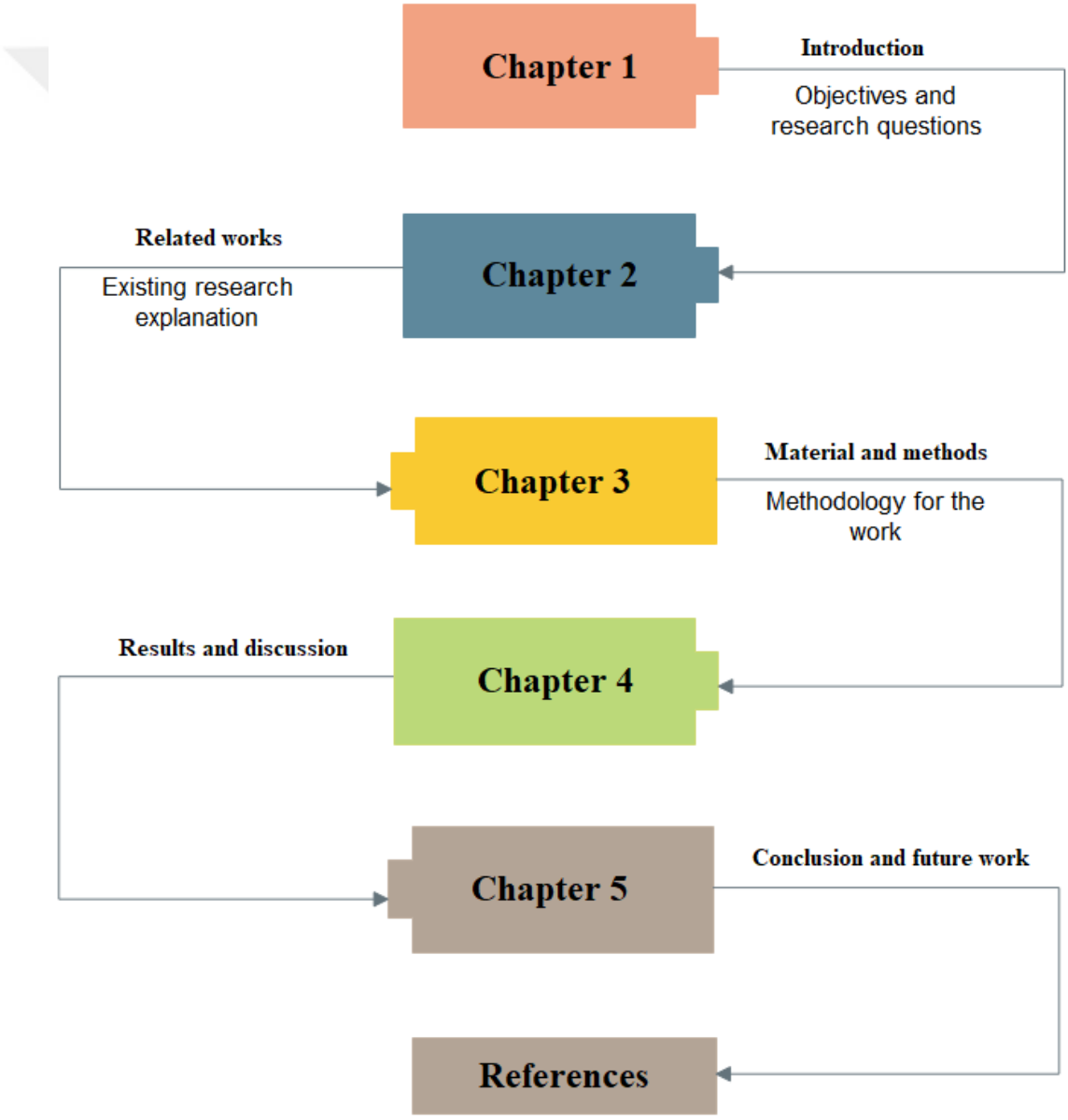


Figure 1.1 Thesis organization

2. LITERATURE REVIEW

In this chapter, we represent a comprehensive overview of the research conducted on the classification of cancer types, specifically focusing on breast, lung, and prostate cancer. We also outline the limitations, advantages, and disadvantages associated with each of these studies.

2.1 Introduction

This chapter introduces an overview about the analysis of cancer diseases and datasets that are used in machine learning techniques . Also, this chapter represnets the related works that were implemented by researchers in this feild. The chapter also introduce a brief camparison between the statistical learning and machine learning and the fainal section introduces the related work to this research and limitations of these researchers and how to solve them using statistical learning.

2.2 Overview of Analysis of Cancer Datasets with AI

Due to the high recurrence and fatality rates, the treatment process is incongruously lengthy and expensive. Accurate early cancer diagnosis and prognosis prediction are crucial for increasing the patient's chance of survival. The accuracy of computational analyses, such as multivariate statistical analysis, is substantially higher than that of empirical forecasts, as a result of advancements in statistics and computer engineering over the years (Abunasser *et al.* 2022).

ML and Deep Learning (DL), and statistical learning, has significantly advanced cancer prediction in clinical research. These classification models, influential in various domains including medicine, continue to undergo research for increased efficiency. The focus is on developing more effective hybrid models by combining diverse methods and algorithms, with a critical emphasis on selecting components that leverage unique characteristics for enhanced precision (Shakeel *et al.* 2020). By constructing such a

mixture and leveraging the benefits of each technique, it is possible to reduce the shortcomings of participatory methods and develop a hybrid model with the fewest defects. Numerous researchers employed ML algorithms as a precursor to statistical methods, including SVM, K-nearest neighbors (KNN), RF and Logistic for cancer detection.

2.3 Importance of the Statistical Techniques in the Analysis of Cancer Datasets

Statistical learning proves highly beneficial in the examination of medical data, specifically in the context of cancer. The outcomes of statistical analyses play a crucial role in making assumptions, implementing preventive measures, and diagnosing cancer. The identification of cancer facilitates the provision of appropriate treatments, ultimately enhancing the chances of recovery for patients. The analysis of cancer registry databases can be divided into two main types: descriptive and analytical. Traditionally, cancer registries have played a fundamental role in providing descriptive data on the occurrence, mortality, prevalence, and survival rates of different cancers, catering to the information needs of the public, scientists, researchers, and policymakers (Who 2019).

2.3.1 What is the goal of statistical learning?

The primary objective of statistical learning theory is to establish a framework for examining the inference problem, which involves acquiring knowledge, generating predictions, making decisions, or constructing models from a given dataset. It addresses the statistical inference challenge of determining a predictive function from data. This theory has resulted in successful applications across various fields (Bousquet *et al.* 2004).

2.4 Comparison of Statistical and Machine Learning Algorithms

In machine learning, practitioners are comfortable with uncertainty, focusing on knowledge extraction from data through techniques like concept formation and pattern detection. Statistical learning, in contrast, emphasizes data analysis and modeling, seeking insights by developing certain models that explain observed patterns. Unlike

machine learners, statistical learners require certainty in their solutions and have a foundation in probability, with machine learning algorithms borrowing from statistical techniques. A statistical model involves utilizing statistical methods to construct a representation of the data and subsequently analyzing it to infer relationships between variables or uncover insights. Conversely, ML employs mathematical or statistical models to gain a comprehensive understanding of the data for predictive purposes (Steyn 2022).

2.5 Related Works About Cancer Datasets

Ganggayah *et al.* (2019) analyzed a breast cancer dataset from the University Malaya Medical Centre, Malaysia (n = 8066, 1993–2016). The dataset had 23 predictor variables and indicated patient survival. Prognostic factors for breast cancer survival were identified using models such as decision tree (accuracy = 79.8%) and random forest (accuracy = 82.7%). Also (Moncada-Torres *et al*. 2021) compared Cox Proportional-Hazards (CPH) regression with machine learning techniques using a dataset of 36,658 non-metastatic breast cancer patients from the Netherlands Cancer Registry. Results showed that machine learning models, especially Extreme Gradient Boosting (XGB), performed at least as well as CPH regression (c-index ~ 0.63), with XGB even surpassing it (c-index ~ 0.73). However (Rohan *et al*. 2019) applied a supervised machine learning approach using the Wisconsin Breast Cancer dataset, which includes 699 instances with 11 features and 10 attributes. Employing the Random Forest ensemble method with AdaBoost, the study achieved significant results in binary classification. During testing, the model demonstrated 98.5714% accuracy, 100% sensitivity, and 96.296% specificity. The calculated Matthews Correlation Coefficient of 0.97 highlights the model's efficacy as a powerful binary classifier for distinguishing between malignant and benign tumor types. (Yifan *et al.* 2021) utilized the Wisconsin female breast cancer tumor dataset to propose a breast cancer classification model, integrating Random Forest and AdaBoost algorithms. In comparison with individual algorithms like SVM, Logistic Regression, KNN, and DT, the ensemble model demonstrated improved prediction accuracy. The highest accuracy achieved was 98.6% with AdaBoost and Random Forest. (Tewari *et al.* 2022) introduced two classifiers, the Naive Bayes (NB) classifier and KNN, for breast

cancer classification. They employed the Breast Cancer Dataset (BCD) from the University of California, Irvine (UCI), containing 699 clinical cases. Through cross-validation, they compared the two classifiers, with KNN achieving the highest accuracy at 97.51% and a lower error rate than the NB classifier, which reached an accuracy of 96.19%. On the other hand, (Sharifmoghadam and Jazayeriy 2019) used an ensemble of Extreme Learning Machines (ELMs) to predict breast cancers, leveraging the Wisconsin Breast Cancer Dataset from the UCI Machine Learning Repository. They introduced an AdaBoost-based algorithm for ensemble learning, incorporating a threshold to reweight misclassified data. The experiment included tuning the number of neurons in the hidden layer, resulting in a high accuracy of 97.14%.

Researchers in (Abunasser *et al.* 2022) proposed a breast cancer detection and classification model with eight cancer classes, using a Kaggle dataset. The evaluation metrics included precision (97.60%), recall (97.60%), and F1-score (97.58%).

Shakeel *et al.* (2020) employs optimized neural and soft computing techniques to address feature set challenges in lung biomedical data. Through bin smoothing normalization, noise is eliminated, and selected features, including minimum repetition and Wolf heuristic, reduce dimensionality. The proposed method, utilizing discrete AdaBoost optimized ensemble learning generalized neural networks, achieves high performance with a minimal error rate of 0.0212 and a prediction rate of 99.48%. On the other hand, (Nageswaran *et al.* 2022) demonstrates accurate lung cancer classification and prediction through machine learning and image processing. The experimental dataset comprises 83 CT scans from 70 distinct patients. Image preprocessing involves using the geometric mean filter, followed by segmentation using the K-means technique. Various machine learning methods, including ANN, KNN, and RF, are employed for classification, resulting in accuracy scores of 95%, 97%, and 98%, respectively.

Gupta *et al.* (2019) implemented multiple ML algorithms to predict the tumor stage of colon cancer, incorporating the Tumor Aggression Score (TAS) as a prognostic factor. Utilized a dataset with information from 4021 patients and evaluated algorithm performance through five-fold cross-validation. Achieved an F-measure of 0.89 when

considering TAS as an attribute alongside standard attributes for TNM stage prediction. Notably, Random Forest exhibited an accuracy of around 84%. Also, (Shafi *et al.* 2020) employing a ML approach, the study analyzes colon cancer data using a random forest classifier and feature selection techniques. The dataset includes 62 cases and 2000 genes from colon cancer patients. The proposed method combines "Mean Decrease Accuracy" and "Mean Decrease Gini" within the Random Forest classifier, reducing high-dimensional data complexity and enhancing calculation speed. The model achieves an impressive overall accuracy of 83.871% using all genes.

Geetha *et al.* (2019) employed risk factors associated with cervical cancer to build a classification model through the Random Forest (RF) method. Tackled dataset imbalance using the synthetic minority oversampling technique (SMOTE) and utilized two feature reduction methods: recursive feature elimination and principal component analysis (PCA). The dataset encompasses 32 risk factors and four outcome variables: Hinselmann, Schiller, Cytology, and Biopsy. The outcomes showcase a notable accuracy rate of 96.06%.

Akai *et al.* (2018) analyzed dynamic computed tomography data from 127 patients with resectable Hepatocellular Carcinoma (HCC). Conducted texture analyses with Random Survival Forests (RSFs) using 96 histogram-based texture features, predicting individual risk for disease-free survival (DFS) and overall survival (OS). Achieved a hazard ratio of 1.06 (95% CI 1.04—1.09) per 1% increase in predicted risk ($P = 8.4 \times 10{-}8$) and 1.74 (95% CI 1.03—2.96; $P = 0.039$) for the presence of vascular invasion. Also, (Rinesh *et al.* 2022) analyzes brain tumor localization in hyperspectral images using k-nearest neighbor and k-means clustering with the firefly algorithm for optimal k values. Multilayer feedforward neural networks label brain areas, resulting in a model with 96.47% accuracy, 96.32% sensitivity, and 98.24% specificity.

Table 2.1 represents the limitations or gaps that we found in the related works according to feature slection techniques, ML algorithms and so on.

Table 2.1 Limitations of related works

RESEARCH	LIMITATIONS
Ganggayah *et al.* (2019)	The limitation of this research they don't use any ensemble learning or stacking learning so the accuracy is very low and they also didn't implement all different ML algorithms
Moncada-Torres *et al.* (2021)	The limitatin of this resarch is the data preprocessing, they don't try all feature selection techniques to selection the best features
Rohan *et al.* (2019)	The limitation of this research summarized in they implemented ensemble but they didn't get ahigh accuracy closer than 100% , I think they didn't used the best single machine learning algorithms for ensemble learning process.
Yifan *et al.* (2021)	This reseach used a very small test samples to evaluate the model, it's not fair to use this number of samples , they must use at least 15% from the total dataset to give the iverall accuracy Also they didn't provide a discussion represent shy the proposed work is better from state of arts.
Tewari *et al.* (2022)	This research work didn't implement the most importnat machine learning algorithms like Adboost, Random forest, ensemble learning also not implemented. These algorithms can get a higher accuracy from the obtained one
Sharifmoghadam and Jazayeriy (2019)	The obtained result for this research is low, it must be higher than this value because the cncer is very importmant disease and it wnt tob be predicted correctly. The degredations of this resaech is not using many ML algorothms and feature selection.
Abunasser *et al.* (2022)	%). Researchers used tranfer learning techniques but they tried one model from them. They had to implement many types from these models and campare between them then selection the best one.
Nageswaran *et al.* (2022)	Researchers in this paper can get higher accuracy when implement stacking or ensemble models but they didn't try.
Gupta *et al.* (2019)	The accuracy is very low because they didn't slect the best features in aperfect way and also the cleaning of the dataset need more attention.
Shafi *et al.* (2020)	The preprocessing steps are not clear in the reasrch and we think this the problem which lead to low accuracy. They also didn't implement the many ML algorithms and compare between them.
Geetha *et al.* (2019)	They implemeted the PCA algorithm to reduce the number of feature without using vaious feature selection techniques. This is a problem in classification process because one feature selection technique may be not suitable for this dataset so, they had to try many feature selection tecnqies and select the best based on the results obtained.
Abdoh *et al.* (2018)	Also this method use only random forest , the ensemble or stackign can be better than using single machine learning algorithm.
Akai *et al.* (2018)	The limitstion of this method is using a very small dataset, it' not suitable to create a model that can be used in hospitals for dignosing patient diseases.
(Rinesh et al. 2022)	The limitation of this method is not using transfer learning algorithms or deep learning because the datset set are images. These algorithm are very effective than machine learning algorithms when using images as a dataset.

As seen in Table 2.1 , there are many limitations for all related work. Our challenges in using statistical learning is finding a suitable dataset that can be used in this learning.

Statistical learning need some features like time if treatment , evnt which indicate the death and other feaures which are important in the learning.

In the domain of medical diagnostics, the robustness of ML models assumes paramount significance, particularly in the context of cancer type classification using diverse modalities such as breast, lung , and prostate data (Lyu *et al.* 2022). The inherent variability in biological signals necessitates models that can generalize well across different patient populations and conditions. Robust ML models can reliably classify cancer types even when faced with subtle variations in the input data, contributing to their effectiveness in real-world clinical settings (Tran et al., 2021). When considering the classification of breast, lung, and prostate cancer, it becomes imperative to leverage distinct datasets for training and evaluation. Each cancer type exhibits unique characteristics and biomarkers, demanding specialized models tailored to the intricacies of each dataset. By employing diverse datasets, the models can learn the nuanced patterns associated with each cancer type, enhancing their ability to discern between different conditions. This approach not only bolsters the accuracy of classification but also ensures that the models are capable of handling the inherent heterogeneity in cancer manifestations. Consequently, the utilization of varied datasets in cancer type classification not only aligns with the principles of robust ML but also reinforces the models' reliability in addressing the multifaceted challenges posed by different cancer types.

We want to fill the gaps from the previous work by implementing statistical leanring because it aids in comprehending the behavior of a system, diminishing uncertainty, and generating meaningful real-world outcomes.

3. MATERIALS AND METHODS

This chapter presents the steps of the methodology used, along with descriptions of the datasets. Figure 3.1 shows the process of the methodology, starting with reading the datasets and then applying preprocessing techniques such as data scaling, cleaning, and splitting. The next step is to implement statistical and machine learning algorithms, followed by evaluating the models for each dataset and discussing the best models.

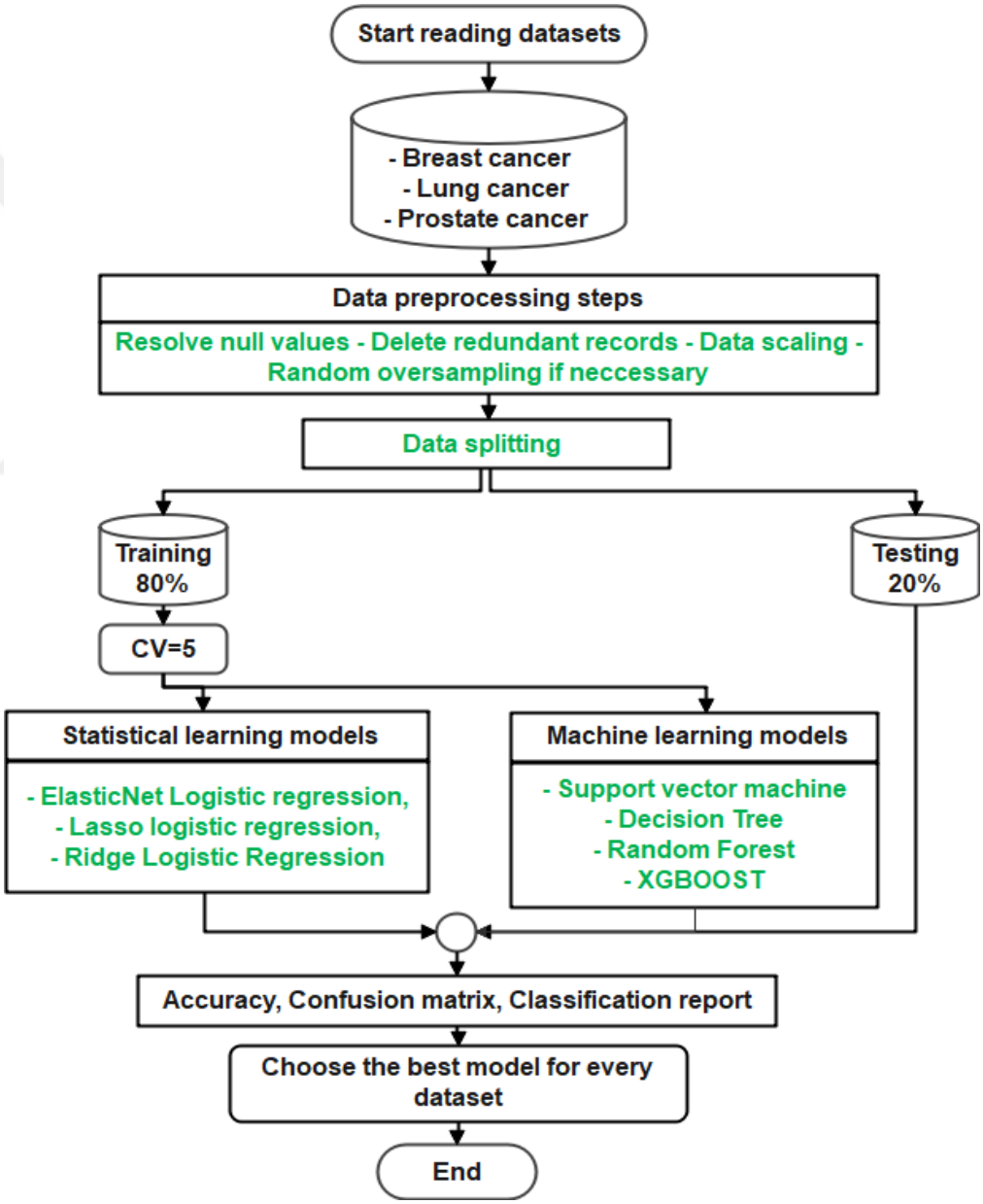


Figure 3.1 Block diagram for the proposed methodology

3.1 Datasets Description

In this part, we will describe the three datasets used in the thesis. They include datasets for breast cancer, lung cancer, and prostate cancer.

3.1.1 Breast cancer dataset

The breast cancer dataset contains 32 features and 569 records, which can be accessed through the refercence (Wolberg 1995). Table 3.1 represent some features and their description from the dataset.

Table 3.1 Some features explanation for the breast cancer dataset

FEATURE	EXPLANATION	VALUES / DATA TYPE
Concave Points	Number of concave portions of the contour	Number
Fractal Dimension	Coastline approximation	Number
Radius	SE of mean distance to perimeter	Number
Texture	SE of std deviation of gray values	Number
Perimeter	SE of the perimeter	Number
Area	SE of the area	Number
Diagnosis	Diagnosis for the patient	M = malignant, B = benign

3.1.2 Lung cancer dataset

The information in this dataset pertains to patients with lung cancer and includes details. The dataset consists of 16 features and 310 records and can view at https://www.kaggle.com/datasets/nancyalaswad90/lung-cancer. Table 3.2 represent the features description for the lung dataset.

Table 3.2 Features descriptions for the lung dataset

FEATURE	EXPLANATION	VALUES / DATA TYPE
Gender	Patient gender	Male, Femal
Age	Patient age	Number
Smoking	Patient smoke or not	Yes/No
Yellow fingers	Color of the finger	Yes, No
Anxiety	Anxiety	Yes, No
Peer_pressure	If exist peeer pressure or not	Yes, No

Table 3.2 Features descriptions for the lung dataset (Continued)

FEATURE	EXPLANATION	VALUES / DATA TYPE
Chronic Disease	If exist chronic disease or not	Yes, No
Fatigue:	If there is fatigue or not	Yes, No
Allergy	If there are any allergy	Yes, No
Alcohol	If the patient drink alcohol	Yes, No
Coughing	If the patient cough	Yes, No
Shortness of Breath	If there is shortness of the breath	Yes, No
Swallowing Difficulty	If the patient swallow in dificult way	Yes, No
Chest Pain	If there is chest pain	Yes, No
Wheezıng	Presence of wheezing	Yes, No
Lung_Cancer	Lung cancer diagnosis	Yes, No

Table 3.3 represent a sample of five frows from the datset withonly five features. The dataset avialbale in research (Hong and Yang 1991).

Table 3.3 Sample five rows from the lung dataset

GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY
M	69	1	2	2
M	74	2	1	1
F	59	1	1	1
M	63	2	2	2
F	63	1	2	1

3.1.3 Prostate cancer dataset

Prostate dataset containing information about 424 patients that can be used to implement a machine learning algorithm and interpret the results. The dataset includes 424 observations and 32 features. The dataset is available in the link: (*https://www.cbioportal.org/study/clinicalData?id=prad_mcspc_mskcc_2020*). Table 3.4 represent the features and their descriptions.

Table 3.4 Some feature descriptions for the prostate dataset

FEATURE	EXPLANATION	VALUES / DATA TYPE
Patient ID	Patient id	Number
Fraction Genome Altered	Proportion of altered genomic material.	Float number
12-245 Part A Consented	Consent for study's Part A	Yes / No
12-245 Part C Consented	Consent for study's Part C	Yes / No
Age at Sample Collection	Age during sample collection	Number
Race Category	Participant's ethnic background	White / other
Androgen Deprivation Therapy (ADT):	Reatment to reduce androgen levels	Number
Biopsy Gleason Grade	Cancer cell aggressiveness score.	Number
Prostate-specific Antigen (PSA)	Prostate-specific protein level	Number
Timing of Metastases	Metastasis occurrence timing	De-novo metastatic, recurrent
Tissue Site	Specific sample collection site	Lymph Node, Prostate, Bone, Other
Race Category	Race of the patient	Text
Sample Class	Class of the sample	Text
Number of Samples Per Patient	Number of samples collected per patient	Numeric
Sample Coverage	Coverage of the sample	Numeric
Sex	Gender of the patient	Text
Time from Sample to CR Resistance	Time from sample to castration resistance (Months)	Numeric
Somatic Status	Somatic status of the sample	Text
SO Comments	Comments related to sample analysis	Text
Survival Status	Survival status of the patient	Text
TMB (Nonsynonymous)	Tumor Mutational Burden (Nonsynonymous)	Numeric
Tumor Purity	Purity of the tumor sample	Numeric
Disease Volume	Extent of disease spread	Low / High

Table 3.5 represent sample five rows from the prostate dataset with only five features' values.

Table 3.5 Sample five rows from the prostate dataset

AGE AT SAMPLE COLLECTION	ANDROGEN DEPRIVATION THERAPY (ADT)	ARCHER PANEL	BIOPSY GLEASON GRADE	CANCER TYPE
55	No	NO	9-10	Prostate Cancer
57	Yes	NO	8	Prostate Cancer
67	No	NO	8	Prostate Cancer
64	Yes	NO	9-10	Prostate Cancer
57	Yes	NO	9-10	Prostate Cancer

3.2 Data Preprocessing

In this section, we will discuss the preprocessing steps utilized in our methodology, such as data cleaning, data scaling, and data splitting.

3.2.1 Data cleaning

Data cleaning is a crucial part of the data preparation process, aiming to ensure the accuracy, completeness, and reliability of data used for analysis or modeling. It involves managing null values and eliminating irrelevant features. Below is a guide on how to address null values and remove unnecessary features:

1) Identify and Handle Null Values: Examine the dataset to locate columns containing null or missing values. Utilize functions such as `isnull()`, `info()`, or `describe()` depending on the programming language or library you are using. We found that there are not values in the datasets.
2) Delete unnecessary features: Identify features that do not contribute to the analysis or modeling objectives. These may include unique identifiers (e.g., IDs), constant features (uniform values for all samples), or features with minimal variance and limited informational value. There were ID columns in the breast and the prostate datasets and we removed them.

3) Delete duplicated records: There are some datasets contain duplicate records that don't contribute to the efficiency of the code so, it's must to delete them. We found 33 redundant records in the lung dataset and we removed them. The other dataset doesn't have redundant records.

3.2.2 Data scaling

Data scaling is a method employed during data preprocessing to ensure that all features in a dataset are on a comparable scale. This is especially significant for machine learning algorithms that can be influenced by the varying magnitudes of features. One widely used technique for data scaling is known as Standard Scaler (Ahsan *et al.* 2021).

With the Standard Scaler, features are transformed to have a mean of zero and a standard deviation of one. This process follows these steps:

- Compute the average and standard deviation for each feature in the dataset.
- Subtract the mean from each feature value.
- Divide the result by the standard deviation to achieve the scaled value.

By applying the Standard Scaler, features become standardized, contributing to improved model performance, particularly when algorithms rely on distance measurements or gradient-based optimization.

3.2.3 Random over sampling

Random oversampling is a method employed to tackle imbalanced datasets, where one class is notably underrepresented. This technique entails replicating instances of the minority class to achieve a more balanced class distribution. Its application enhances the effectiveness of machine learning models, particularly those susceptible to issues arising from class imbalance (Geetha *et al.* 2019).

By increasing the minority class instances, the classifier can better learn the class patterns, leading to more accurate predictions. However, random oversampling may cause overfitting, where the model memorizes the data.

We used this technique in the breast and lung cancer dataset because they are no balanced. The breast cancer diagnosis dataset initially had 0.62% and 0.37% samples for the first and second classes of the output column, respectively. To address the class imbalance, we decided to balance this dataset. Similarly, the lung dataset had 86% and 13% samples for the first and second classes, respectively, prompting us to also balance this dataset. Table 3.6 represent the number of instances before and after balancing using this technique.

Table 3.6 Number of samples of the dataset before and after oversampling

DATASET	**NUMBER OF SAMPLES**	
	BEFORE BALANCING	**AFTER BALANCING**
Breast cancer	569	714
Lung cancer	310	476
Prostate cancer	No need to balance	

As seen in Table 3.6, initially, the Breast cancer dataset had 569 samples, and the Lung cancer dataset had 310 samples, indicating class imbalance. To address this, random oversampling was used, increasing the samples in each class. After balancing, the Breast cancer dataset had 714 samples, and the Lung cancer dataset had 476 samples, resulting in improved representation of the minority class and enhancing model training and evaluation. The prostate dataset doesn't need to be balances because it's already need to balancing.

3.2.4 Data splitting

In this thesis focused on cancer analysis, a vital aspect of data preparation is data splitting. Each cancer dataset (breast, lung, and prostate) undergoes the process of splitting into two groups: one for training (80%) and the other for testing (20%). This strategy ensures effective training and evaluation of machine learning models. The training set imparts patterns, relationships, and features to the models using the data, while the testing set acts

as an independent measure to assess how well the models perform on new and unseen data. By adopting the 80:20 ratio, the thesis achieves a balance between learning from a significant portion of data and reserving sufficient unseen data, preventing overfitting and ensuring the analysis remains dependable and robust for cancer detection and classification (Picard and Berk 1990). Table 3.7 represent a brief description for the splitting process for the used three datasets.

Table 3.7 Data splitting for the three datasets

DATASET	TRAINING SAMPLES 80%	TESTING SAMPLES 20%	TOTAL
Breast cancer	571	143	714
Lung cancer	380	96	476
Prostate cancer	340	84	424

As shown in Table 3.7, for Breast cancer, there are 571 training samples and 143 testing samples, totaling 714 samples. The Lung cancer dataset contains 380 training samples and 96 testing samples, making a total of 476 samples. The Prostate cancer dataset includes 340 training samples and 84 testing samples, totaling 424 samples. This data splitting approach ensures models are trained on a portion of the data while preserving unseen samples for unbiased evaluation. These datasets are crucial for advancing cancer research and developing accurate models for cancer detection and classification.

3.2.5 Cross validation

In this thesis, Cross Validation with k=5 is used to assess machine learning models and statistical learning models on breast, lung, and prostate cancer data. The training sets are divided into five subsets (folds), rotated as training and validation data across five iterations. Performance metrics are computed for each fold, and the final evaluation is the average across all iterations. Cross Validation enhances the models' reliability, reducing overfitting risk by testing on diverse data partitions. It ensures the models' effectiveness and generalization, making them applicable to new and unseen cancer data with reasonable accuracy (Ojala and Garriga 2011). Figure 3.2 represent the process of cross validation with using cross validation value =5.

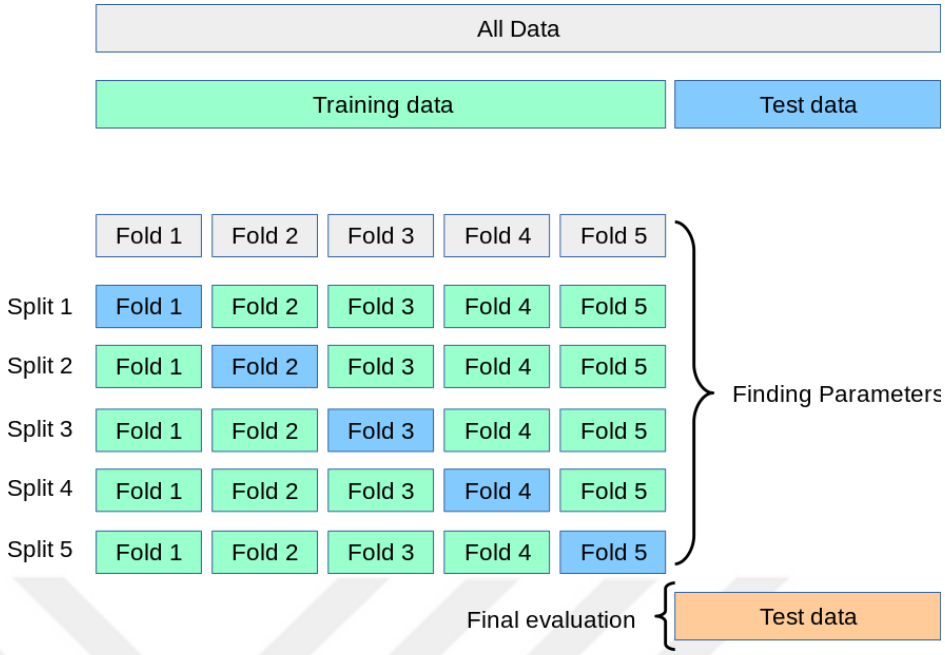


Figure 3.2 Illustration of cross validation process with CV=5 (Ojala and Garriga 2011)

As depicted in Figure 3.2, the training data is split into five segments. During each step, four segments are utilized for training, and the remaining segment is dedicated to the validation process.

3.3 Statistical Learning Algorithms

In the domain of machine learning and data analysis, statistical learning algorithms play a crucial role, enabling researchers to extract valuable insights and make precise predictions from complex datasets. Among these algorithms, Lasso logistic regression, and ridge logistic regression and Elastic Net logistic regression have emerged as indispensable tools, especially when confronted with high-dimensional and correlated data that may pose limitations to traditional logistic regression.

3.3.1 Lasso logistic regression

Lasso logistic regression is a ML algorithm designed for binary classification tasks. It is a regularized version of logistic regression that utilizes L1 regularization for feature selection. In logistic regression, the aim is to find coefficients that minimize the log-loss

function between predicted probabilities and binary labels (Abinash and Vasudevan 2019).

In Lasso logistic regression, an L1 regularization term is added to the objective function, penalizing coefficients based on their magnitudes. This induces sparsity in the solution, shrinking many coefficients to zero and effectively performing feature selection. The objective function can be expressed as minimizing the log-loss plus λ times the L1 norm of the coefficient vector.

The objective function of Lasso logistic regression can be written as in Equation (3.1).

$$minimize\ \frac{1}{N} * \ sum\big(y\ *\ log(p) + \ (1-y) * \ log(1-p)\big) + \ \lambda\ *\ \big||\beta|\big| \qquad (3.1)$$

Where N is the number of data points, y is the binary label of each data point, p is the predicted probability of the positive class, β is the vector of coefficients, and λ is the regularization parameter that controls the strength of the regularization.

Increasing λ leads to more coefficients being shrunk to zero, thereby removing corresponding features from the model. An iterative optimization algorithm, like coordinate descent, finds the optimal coefficient values.

Lasso logistic regression offers several advantages:

- Feature Selection: It selects relevant features, improving model performance and reducing overfitting.
- Interpretability: Selected features are easily interpretable, enhancing model understanding.
- Regularization: L1 regularization prevents overfitting by diminishing irrelevant feature impact.
- Efficiency: Lasso logistic regression efficiently handles large datasets.

- Flexibility: It can handle both linear and nonlinear relationships between inputs and the target variable.

For the implementation of Lasso logistic regression, the grid search technique choose the best parameters which are 'C': 1.0, 'penalty': 'l1', 'solver': 'liblinear'.

3.3.2 Ridge logistic regression

Ridge logistic regression is a binary classification machine learning algorithm. It acts as a regularized version of logistic regression, employing L2 regularization to combat overfitting and enhance model generalization. The primary objective in logistic regression is to find coefficient values that minimize the log-loss function between predicted probabilities and binary labels (Hoerl and Kennard 2000).

In Ridge logistic regression, an L2 regularization term is added to the objective function, penalizing coefficients based on their magnitudes. This encourages small coefficients and a smoother model, reducing the impact of irrelevant features and limiting coefficient magnitudes, which improves generalization performance.

The objective function as seen in Equation (3.2) for Ridge logistic regression is a combination of the log-loss and the L2 norm of the coefficient vector. As λ (the regularization parameter) increases, the regularization term's impact becomes more pronounced, leading to reduced coefficient magnitudes and better generalization on new data.

$$minimize \frac{1}{N} * sum\big(y * log(p) + (1-y) * log(1-p)\big) + \frac{\lambda}{2} * \big|\big|\beta\big|\big|2^2 \qquad (3.2)$$

Where N is the number of data points, y is the binary label of each data point, p is the predicted probability of the positive class, β is the vector of coefficients, and λ is the regularization parameter that controls the strength of the regularization.

Optimal coefficient values are found using iterative optimization algorithms like gradient descent, which updates coefficients in the direction of the negative gradient of the objective function.

Advantages of Ridge logistic regression:

- Regularization: L2 regularization prevents overfitting by reducing the impact of irrelevant features and limiting coefficient magnitudes.
- Improved Generalization: By mitigating overfitting, Ridge logistic regression enhances model generalization on new data.
- Stability: Ridge logistic regression exhibits more stability than standard logistic regression, as small changes in input data do not cause substantial coefficient changes.
- Flexibility: Ridge logistic regression can handle both linear and nonlinear relationships between features and the target variable.

For the implementation of Ridge logistic regression, the best parameter for this model is with alpha=1.

3.3.3 ElasticNet logistic regression

Elastic net logistic regression is a regularized variant of logistic regression that merges L1 and L2 regularization techniques to achieve feature selection and counteract overfitting. It balances the strengths of both types of regularization, providing a robust and flexible approach to modeling (Zou and Hastie 2005).

The objective function of elastic net logistic regression can be written as in Equation (3.3).

$$minimize \frac{1}{N} * sum\big(y * log(p) + (1-y) * log(1-p)\big) + \lambda 1 * \big|\big|\beta\big|\big| + \frac{\lambda 2}{2} * \big|\big|\beta\big|\big|2^2 \quad (3.3)$$

In the provided context, N stands for the quantity of data points, y indicates the binary label assigned to each data point, p represents the predicted probability of the positive class, β is the vector of coefficients, and λ1 and λ2 are the regularization parameters that control the intensities of L1 and L2 regularization, respectively.

In summary, the elastic net logistic regression process involves data collection and preprocessing, data splitting into training and test sets, defining the model with both L1 and L2 regularization, training the model using an iterative optimization algorithm, and testing the model's performance on the test set. The objective function of elastic net logistic regression includes terms for both L1 and L2 regularization, with hyperparameters λ1 and λ2 controlling the strengths of each regularization type (Hastie *et al.* 2010).

Advantages of Elastic Net Logistic Regression:

- Feature selection: Shrinks coefficients to zero, useful for high-dimensional data with irrelevant features, enhancing interpretability and reducing overfitting.
- Robustness: More robust to multicollinearity than L1 regularization alone, as the L2 term addresses multicollinearity by shrinking correlated coefficients.
- Flexibility: Offers a flexible trade-off between L1 and L2 regularization, adapting to diverse data types and balancing feature selection and model complexity.
- Better performance: Often outperforms L1 or L2 regularization alone, especially with high-dimensional or highly correlated data, and handles multiple classes.
- Easy implementation: Accessible to researchers and practitioners with varying expertise, implemented using standard ML libraries and software.

The parameters used for the implementation of the Elastic Net logistic regression are penalty='elasticnet', l1_ratios=[0.1, 0.5, 0.9], solver='saga'

3.4 Machine Learning Algorithms

This section illustrates the proposed ML algorithms for classification of three cancer types. These algorithms are the SVC, DT, Random Forest and XGBOOST.

3.4.1 Support vector machine

The SVC is a potent ML algorithm designed for classification tasks, falling within the broader category of SVM. Its main objective is to discover a hyperplane that efficiently distinguishes data points of various classes within the feature space. In binary classification, this hyperplane serves as a decision boundary that maximizes the margin between classes, with essential support vectors delineating it (Noble 2006).

SVC proves highly valuable for handling complex and high-dimensional datasets, accommodating both linearly and non-linearly separable data through various kernel functions like linear, polynomial, Radial Basis Function (RBF), and sigmoid kernels. The kernel trick enables the algorithm to transform the feature space and efficiently identify non-linear decision boundaries.

Its efficacy lies in handling high-dimensional data with relatively few samples, making it well-suited for diverse applications, including image recognition, text categorization, and bioinformatics.

The parameters used for the implementation of SVC re {'C': 10, 'gamma': 0.01} which are obtained by the grid search technique.

3.4.2 Decision tree

The DT is a well-known and interpretable ML algorithms utilized for classification tasks. It belongs to the supervised learning category and constructs a tree-like model, with internal nodes making decisions based on specific features and leaf nodes providing final predictions (Myles *et al.* 2004).

The algorithm recursively divides the data into subsets by considering different feature values. The objective is to identify the best splits at each node, optimizing the separation of target classes or reducing regression variance. This procedure persists until a stopping criterion is satisfied, which could be defined by parameters like maximum depth or a minimum number of data points per leaf.

Decision Trees come with several benefits, including easy interpretation and visualization due to their simple tree structure. They handle both numerical and categorical features, withstand outliers and missing data, and effectively capture non-linear relationships between features and the target variable.

For the implementation process, the best parameters obtained by the grid search are 'criterion': 'gini', 'max_depth': 17, 'min_samples_leaf': 1, 'min_samples_split': 5, 'splitter': 'random'.

3.4.3 Random forest

RF is a widely adopted and potent machine learning ensemble algorithm, excelling in both classification and regression tasks. It extends the DT algorithm by combining predictions from multiple individual trees, leading to more accurate and robust results (Athey *et al.* 2019).

The algorithm generates numerous DTs, each trained on random subsets of data and features. Independence among these trees reduces the risk of overfitting and enhances generalization to new, unseen data. The RF's ultimate prediction is derived by combining individual tree predictions through majority voting (for classification) or averaging (for regression).

RF offers various advantages, including high accuracy, resilience to noisy data, and the ability to handle large and high-dimensional datasets. Compared to a single Decision Tree, it is less prone to overfitting, and the diversity among individual trees contributes to a more dependable and steadier model.

For the implementation, the RF model used 'entropy' as the criterion, had a maximum depth of 11, and used 'auto' to determine the number of features considered for each split. The minimum samples required for a leaf node was 2, and the minimum samples required to split an internal node was 3. It consisted of 130 decision trees (estimators) in the ensemble for accurate and robust predictions.

3.4.4 XGboost

Xgboost, which stands for Extreme Gradient Boosting, represents a sophisticated and remarkably powerful machine learning algorithm that falls under the gradient boosting umbrella. Known for its advanced capabilities, it finds widespread application in tasks involving both classification and regression. Notably, XGBOOST has earned substantial popularity, proving its prowess in various domains, including data science competitions and real-world applications. Its effectiveness in enhancing predictive accuracy and handling complex datasets has contributed to its status as a preferred choice among machine learning practitioners (Chen *et al.* 2018).

Functioning as an ensemble learning technique, XGBOOST combines predictions from multiple weak learners, usually DTs, to create a more precise and robust model. The algorithm adds trees to the ensemble sequentially, each focusing on reducing the errors of its predecessor, thereby continuously refining its predictions.

XGBoost's key strengths lie in its efficiency, scalability, and ability to handle large and high-dimensional datasets. By incorporating various regularization techniques, it guards against overfitting, while its capability to manage missing data makes it suitable for real-world datasets with incomplete information.

The essential XGBOOST parameters used were 'learning_rate' (0.5), 'max_depth' (5), and 'n_estimators' (180). These control model complexity, learning rate, and boosting rounds. Some other parameters like 'missing' were set to 'nan', and 'enable_categorical' was set to False. XGBOOST is an algorithm that optimizes the boosting process, combining weak learners to create a powerful ensemble model for classification and regression tasks.

3.5 Evaluation Metrics

In this section we describe the evaluations metrics used in the thesis for calculating the performance of the statistical and ML algorithms. These metrics are accuracy, confusion matrix and classification report.

3.5.1 Confusion matrix

A commonly used instrument for assessing the performance of a classification model is a confusion matrix. It furnishes a comprehensive summary of the model's predictions in relation to the actual labels. Represented as a square table, the matrix uses rows and columns to depict the predicted and actual class labels, respectively (Hossin and bin Sulaiman 2015).

The confusion matrix has four essential elements. True Positive (TP) is for correctly identified positives, False Positive (FP) for incorrectly identified positives, True Negative (TN) for correctly identified negatives, and False Negative (FN) for incorrectly identified negatives. These elements assess how well a classification model performs by detailing correct and incorrect predictions for positive and negative classes.

3.5.2 Classification report

A classification report provides a detailed evaluation of the model's effectiveness for different classes in the dataset, presenting various evaluation metrics for each class. The report serves as a valuable tool for understanding the model's performance on a per-class basis and helps identify areas that may require improvement (Hossin and bin Sulaiman, 2015).

Various evaluation metrics can be derived from the classification, including:

Accuracy: Proportion of correct predictions (TP and TN) out of the total instances as in Equation (3.4).

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \quad (3.4)$$

Precision: Proportion of true positive predictions (TP) out of all positive predictions (TP and FP), indicating the model's ability to avoid false positives as in Equation (3.5).

$$Precision = \frac{TP}{TP+FP} \quad (3.5)$$

Recall (Also known as Sensitivity or True Positive Rate), Recall is the proportion of true positive predictions (TP) relative to the total number of actual positive instances (TP and FN). It illustrates the model's capacity to correctly identify all positive instances as in Equation (3.6).

$$Recall = \frac{TP}{TP+FN} \quad (3.6)$$

F1 Score: The harmonic mean of precision and recall provides a balanced metric, especially valuable for addressing imbalanced class distributions. As in Equation (3.7).

$$F1\ score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (3.7)$$

3.5.3 Receiver operating characteristic curve

The ROC (Receiver Operating Characteristic) curve illustrates a binary classification model's performance at varying decision thresholds, plotting True Positive Rate against False Positive Rate. The Area Under the Curve (AUC) condenses this into a single metric, ranging from 0 to 1. A higher AUC signifies better model discrimination, with 1 being ideal (Bhandari 2020).

4. RESULTS AND DISCUSSION

In this chapter, we will present the results from three cancer datasets: breast cancer, lung cancer, and prostate cancer. Using advanced statistical and machine learning algorithms, we aimed to gain insights into each cancer type's characteristics, uncover meaningful patterns, and contribute to cancer research advancements for better diagnosis and treatment approaches.

4.1 Results of Breast Cancer Dataset

This section describes the result of statistical and machine learning algorithms used for the breast cancer dataset.

4.1.1 Results of statistical learning algorithms

During this research, we employed advanced statistical learning algorithms like lasso, ridge, and elastic net logistic regression. These potent algorithms were utilized to analyze the cancer datasets, which encompassed breast cancer, lung cancer, and prostate cancer. Table 4.1 presents the outcomes of statistical learning algorithms applied to the breast cancer dataset.

Table 4.1 Results of the statistical learning algorithms for breast cancer

ALGORITHM	ACCURACY	PRECISION	RECALL	F1-SCORE
Lasso logistic regression	97.2	97	97	97
Ridge logistic regression	95.8	96	96	96
Elastic Net logistic regression	96.5	96.5	96.5	96.2

According to Table 4.1, Lasso logistic regression outperformed the others with an accuracy of 97.2%, balanced precision and recall at 97%, and an F1-score of 97%. Ridge logistic regression achieved an accuracy of 95.8% with balanced precision, recall, and F1-score at 96%. Elastic Net logistic regression, positioned between Lasso and Ridge, showed an accuracy of 96.5%, precision and recall both at 96.5%, and an F1-score of 96.2%. These findings highlight the nuanced trade-offs and strengths of each algorithm,

guiding practitioners in selecting an optimal model based on specific objectives and requirements for breast cancer classification. Figure 4.1 represent the confusion matrices for using the statistical learning algorithms.

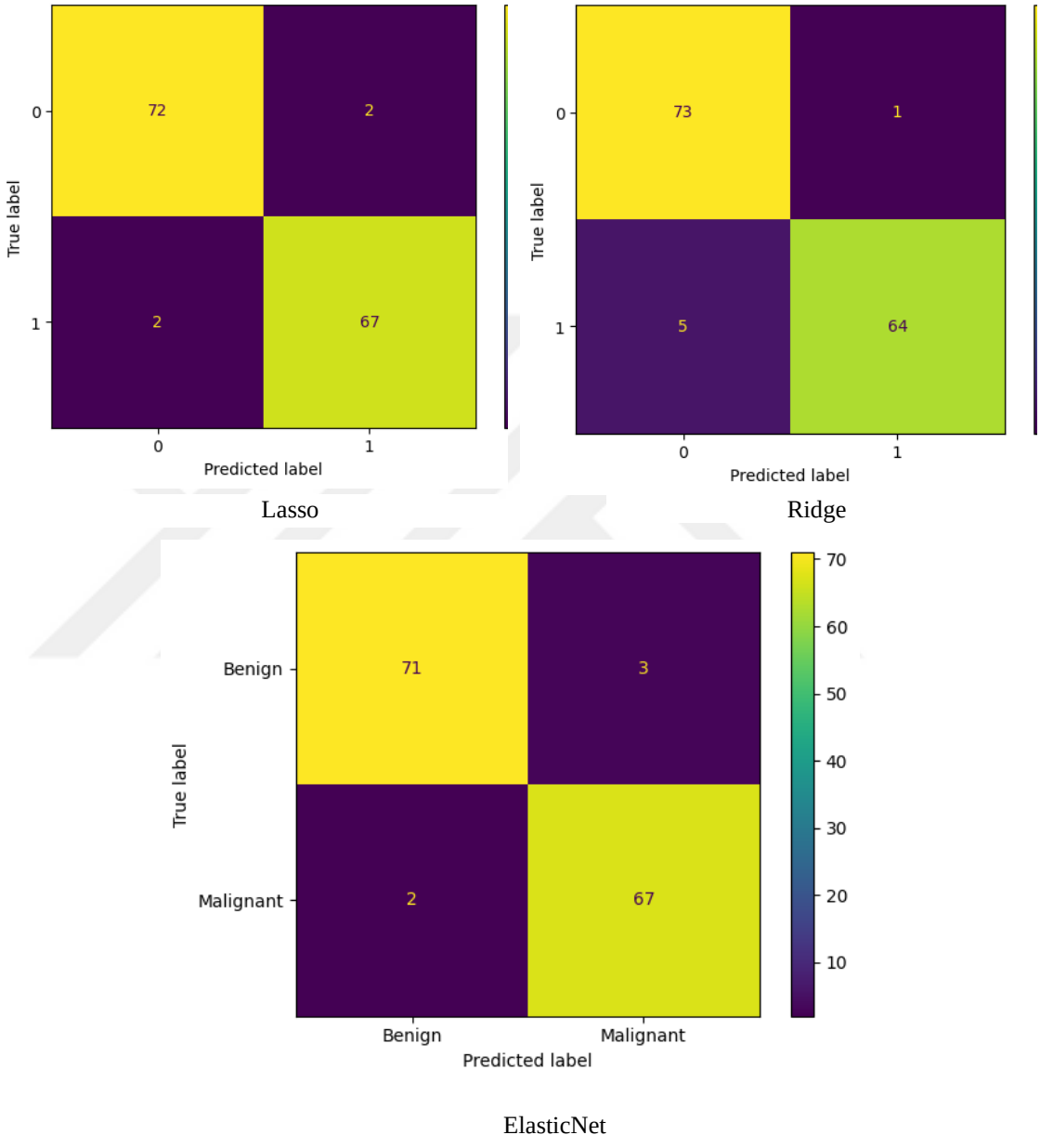
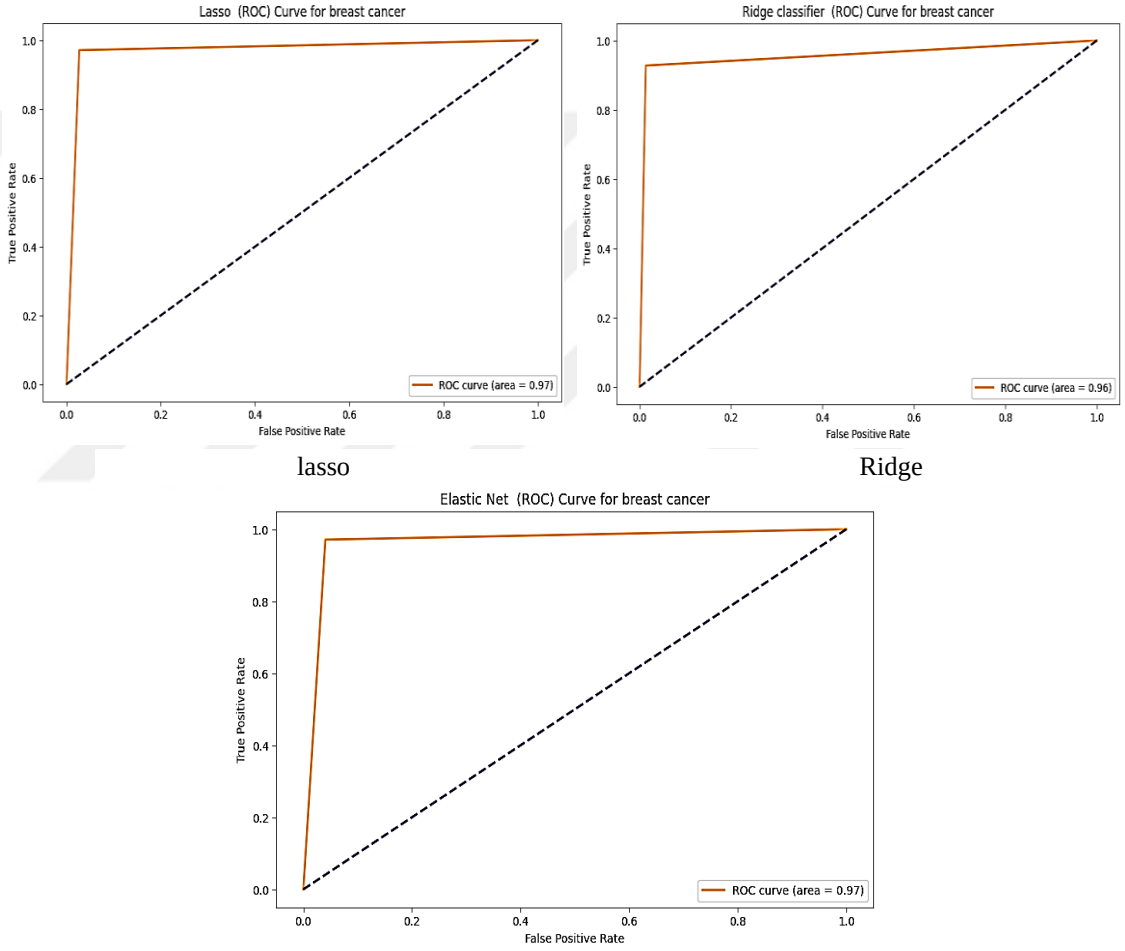


Figure 4.1 Confusion matrices for statistical algorithms on breast cancer dataset

As seen in Figure 4.1, For the Lasso model, there were 72 instances accurately identified as positive (true positives), 2 instances incorrectly identified as positive (false positives), 2 instances incorrectly identified as negative (false negatives), and 67 instances accurately

identified as negative (true negatives). The Ridge model yielded 73 true positives, 1 false positive, 5 false negatives, and 65 true negatives. Lastly, the ElasticNet model demonstrated 71 true positives, 3 false positives, 2 false negatives, and 67 true negatives.

In Figure 4.2, the ROC curves and the associated AUC scores offer a thorough assessment of the breast cancer classification performance across the Lasso, Ridge, and ElasticNet models.



lasso Ridge

ElasticNet

Figure 4.2 AUC scores for the breast cancer using statistical algorithms

As in Figure 4.2, The Lasso model demonstrates a high discriminative ability with an AUC of 0.97, indicating a 97% probability that the model will correctly distinguish between benign and malignant cases. Similarly, the Ridge model exhibits strong

performance, achieving an AUC of 0.96. The ElasticNet model also performs well with an AUC of 0.97. These AUC scores, approaching 1.0, suggest robust and accurate discrimination capabilities, reinforcing the reliability of these models in distinguishing between benign and malignant breast cancer cases.

4.1.2 Results of ML algorithms

Within this section, we showcase the performance metrics derived from utilizing ML algorithms. These metrics provide valuable insights into the efficacy and proficiency of these algorithms when applied to their respective tasks. Table 4.2 showcases the results of different machine learning algorithms applied to the breast cancer dataset.

Table 4.2 Results of ML algorithms for breast cancer dataset

ALGORITHM	ACCURACY	PRECISION	RECALL	F1-SCORE
SVM	97.9	97.95	97.82	97.88
Decision Tree	94.4	94.30	94.48	94.38
Random Forest	93	93.06	93.1	93.01
XGBoost	98.6	98.59	98.59	98.59

As shown in Table 4.2, SVM stands out with a high accuracy of 97.9%, indicating its capability to properly classify instances into benign or malignant categories. The precision, recall, and F1-score for SVM are also impressive, all above 97%, suggesting a balanced performance in terms of correctly identifying positive cases, minimizing false positives, and capturing true positive instances. The Decision Tree algorithm follows with a slightly lower but still commendable accuracy of 94.4%, demonstrating good overall classification performance. Random Forest, an ensemble method, achieves an accuracy of 93%, while XGBoost outperforms all algorithms with an accuracy of 98.6%. Figure 4.3 represent the confusion matrices for using the ML algorithms for breast cancer dataset.

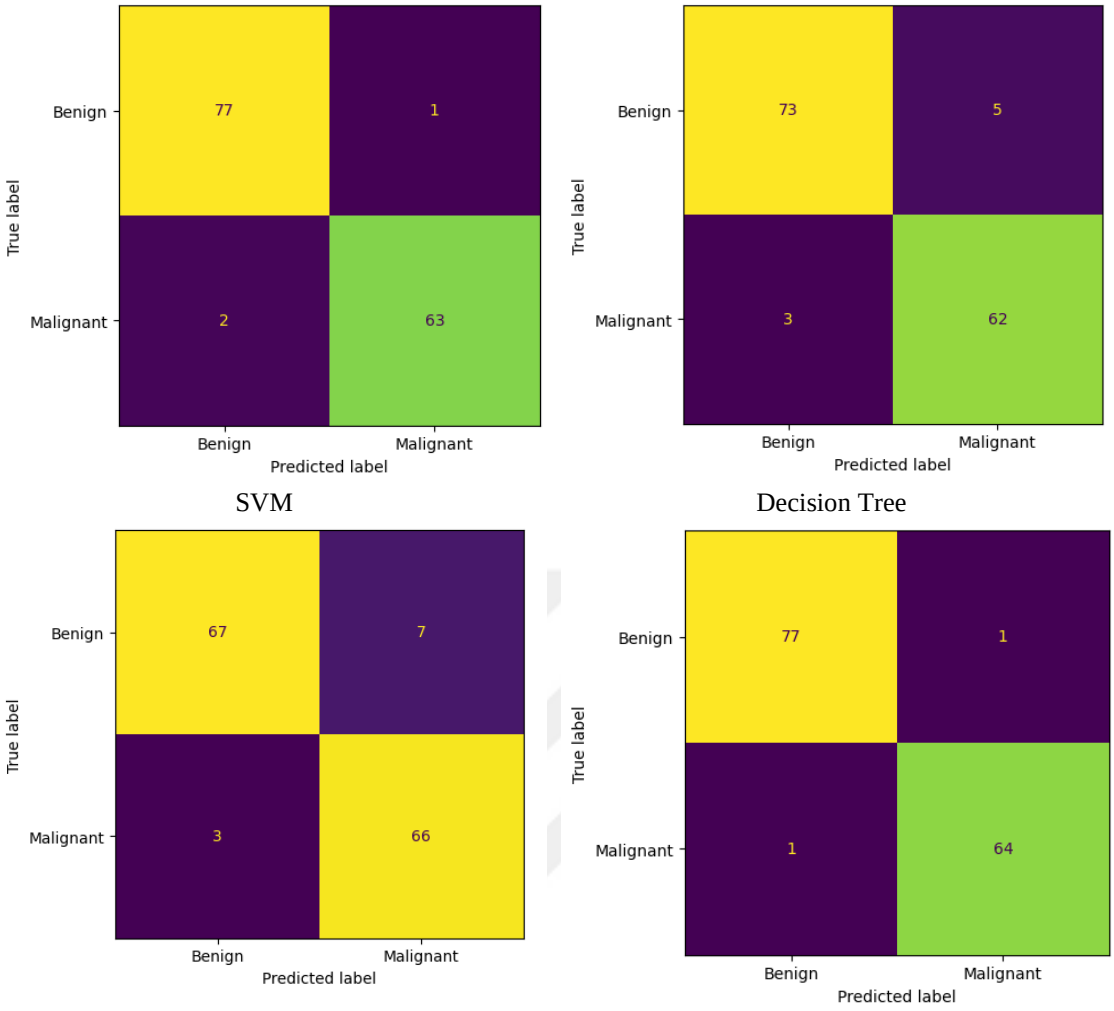
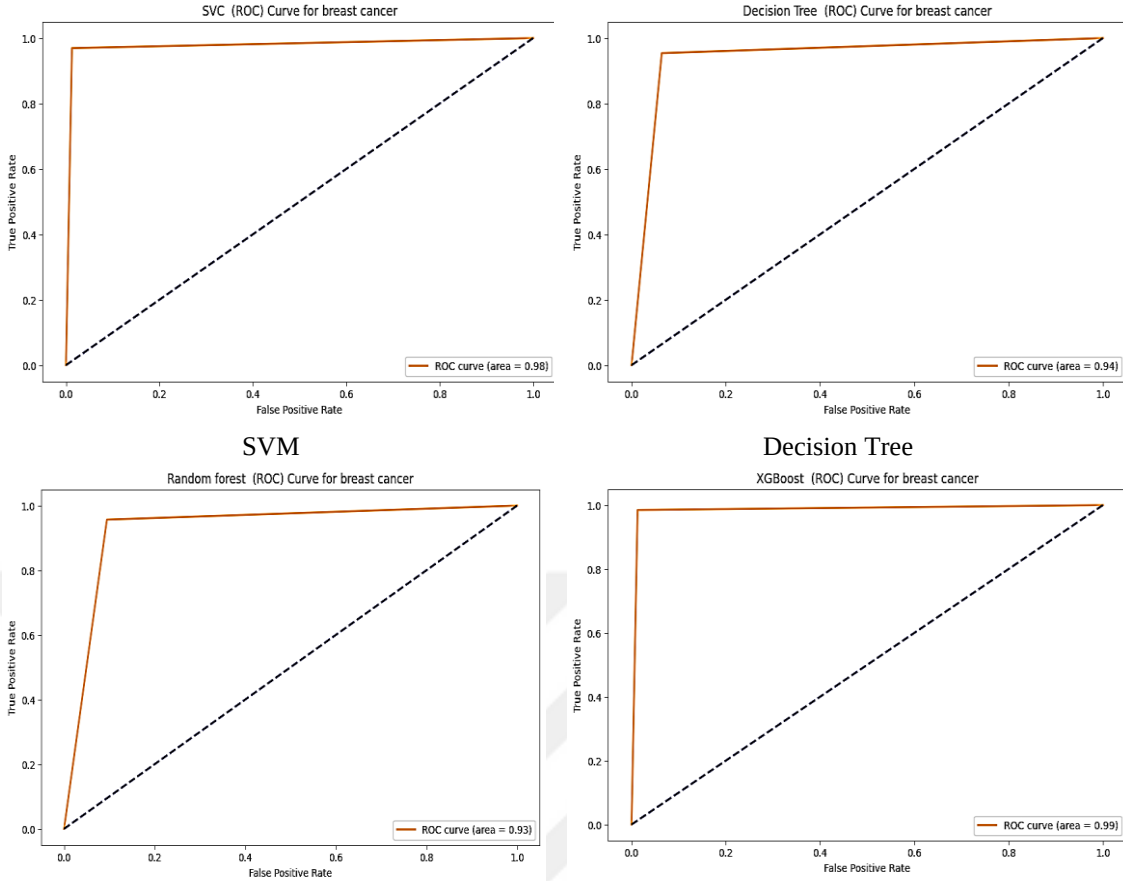


Figure 4.3 Confusion matrices for ML algorithms for breast cancer

The confusion matrices in Figure 4.3 show that SVM model reveals 77 true benign cases, 1 false positive, 2 false negatives, and 63 true malignant cases. The Decision Tree model exhibits 73 true benign cases, 5 false positives, 3 false negatives, and 62 true malignant cases. The Random Forest model's confusion matrix shows 67 true benign cases, 7 false positives, 3 false negatives, and 66 true malignant cases. Finally, the XGBoost model achieves 77 true benign cases, 1 false positive, 1 false negative, and 64 true malignant cases. Figure 4.4 represents The ROC curves and associated AUC scores provide a comprehensive evaluation of the breast cancer classification performance for different machine learning models.



SVM Decision Tree

Random Forest XGBoost

Figure 4.4 AUC scores for the ML proposed algorithms

According to Figure 4.4, the SVM demonstrates excellent discriminative ability with an AUC of 0.98, indicating a 98% probability that the model will accurately distinguish between benign and malignant cases. The Decision Tree model exhibits a respectable AUC of 0.94, showcasing its ability to differentiate between the two classes. The Random Forest model follows closely with an AUC of 0.93, reflecting a strong discriminatory performance. Notably, XGBoost outperforms all others with an impressive AUC of 0.99, suggesting near-perfect sensitivity and specificity. These AUC scores reinforce the robustness and effectiveness of the machine learning models in distinguishing between benign and malignant breast cancer cases, with XGBoost standing out as particularly promising in this regard.

4.2 Results of Lung Cancer Dataset

In this section, we showcase the results of applying statistical and machine learning algorithms to the dataset on lung cancer.

4.2.1 Results of statistical learning algorithms

This section introduces the results of using lasso, ridge and elastic net logistic regression algorithms. Table 4.3 displays the evaluation results of statistical algorithms applied to the lung cancer dataset.

Table 4.3 Results metrics for Statistical algorithms for lung cancer

ALGORITHM	ACCURACY	PRECISION	RECALL	F1-SCORE
Lasso Logistic Regression	96.87	97.06	96.87	96.87
Ridge Logistic Regression	93.75	94.37	93.13	93.57
ElasticNet Logistic Regression	95.8	96.55	95.24	95.71

As in Table 4.3, Lasso logistic regression achieved a high accuracy of 96.87%, with precision, recall, and F1-Score all at 96.87%. Ridge logistic regression followed closely with an accuracy of 93.75% and well-balanced precision (94.37%) and recall (93.13%), yielding an F1-Score of 93.57%. Elastic Net logistic regression demonstrated intermediary performance with an accuracy of 95.8%, precision of 96.55%, recall of 95.24%, and an F1-Score of 95.71%. While Lasso excelled in precision and recall, the choice among these models depends on specific application requirements, considering the nuanced trade-offs in their performance. Figure 4.5 represents the confusion matrices for lung cancer classification using Lasso, Ridge, and ElasticNet models provide a detailed insight into the models' performance.

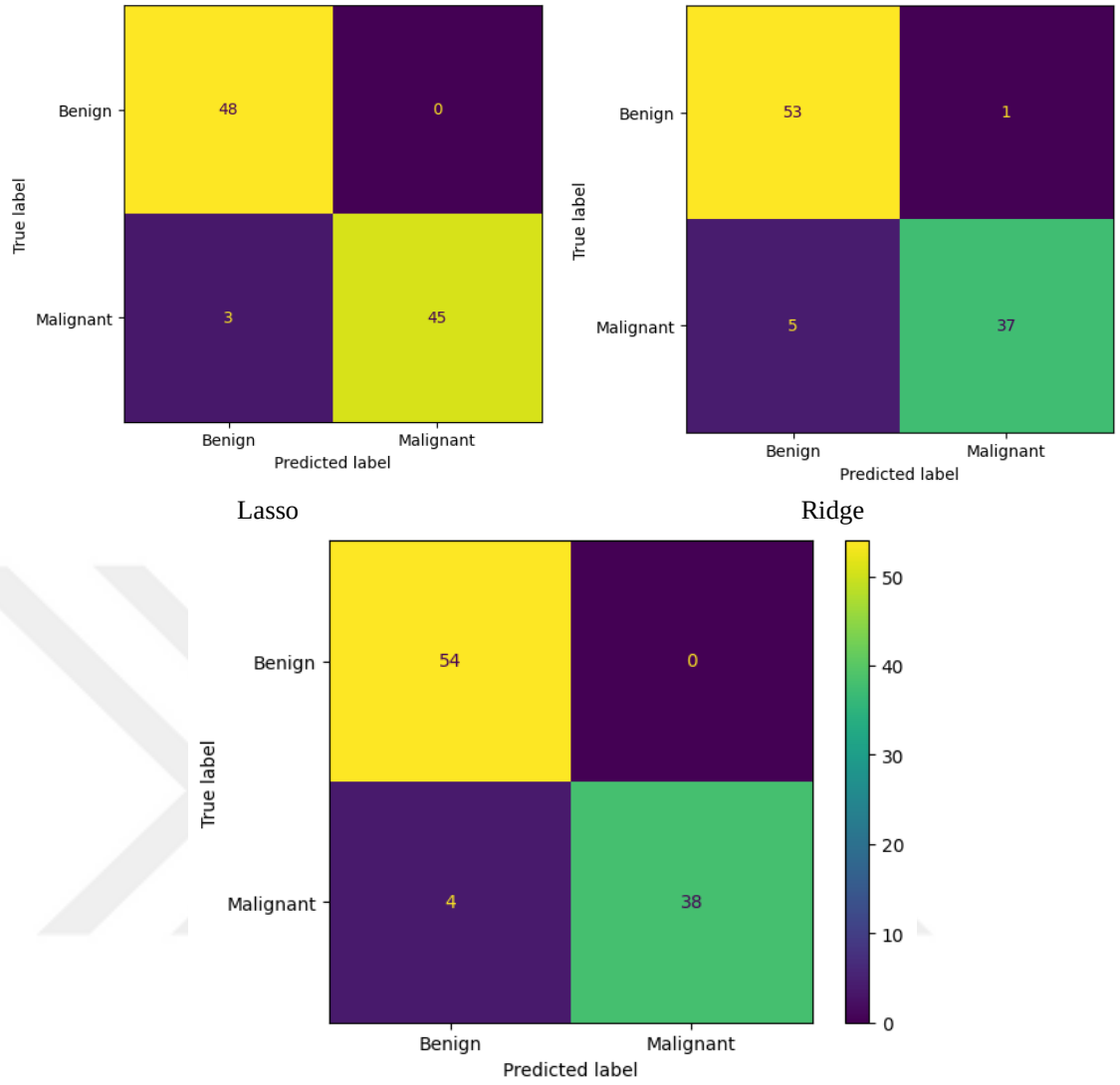
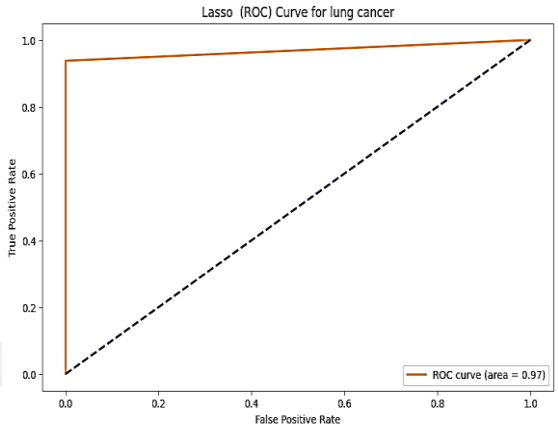


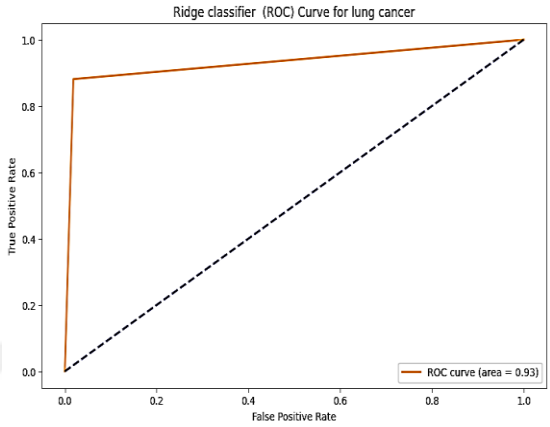
Figure 4.5 Confusion matrices for statistical learning algorithms for lung dataset

According to Figure 4.5, for the Lasso model, it correctly classified 48 cases as benign and 45 cases as malignant, with no false positives or false negatives. The Ridge model exhibited a slightly higher false positive rate, misclassifying one benign case, but still demonstrated good overall performance with 53 true negatives and 37 true positives. The ElasticNet model showcased a robust performance, correctly classifying 54 benign cases and 38 malignant cases, with no false positives or false negatives. These confusion matrices highlight the models' ability to discriminate between benign and malignant cases, with variations in their error types. The absence of false positives and false negatives in the Lasso and ElasticNet models indicates a high precision and recall, while

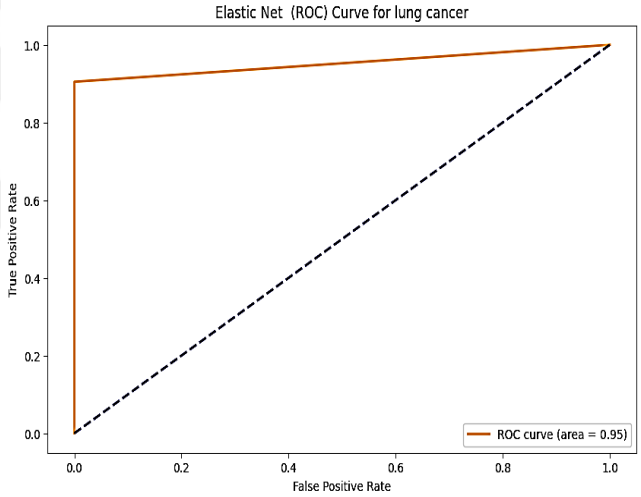
the Ridge model, with minimal misclassifications, demonstrates a balanced performance in the classification task. Figure 4.6 show the ROC AUC scores for the statistical learning algorithms.



Lasso



Ridge



ElasticNet

Figure 4.6 ROC AUC scores for the statistical algorithms for lung cancer dataset

According to Figure 4.6, the Lasso model demonstrated a highly effective discriminatory ability, as evidenced by its ROC AUC score of 0.97. The Ridge model exhibited a slightly lower but still commendable ROC AUC area of 0.93, highlighting its ability to distinguish between benign and malignant cases. Similarly, the ElasticNet model yielded a ROC AUC area of 0.95, reflecting its robust discriminatory power. These AUC scores suggest that all three models have strong predictive capabilities in distinguishing between benign

and malignant lung cancer cases, with the Lasso model achieving the highest discriminatory performance among them.

4.2.2 Results of machine learning algorithms

In this section, we present the results of applying ML algorithms to analyze the lung cancer dataset. Table 4.4 gives a comprehensive overview of the outcomes achieved by various algorithms, including SVM, Decision Tree, Random Forest, and XGBoost.

Table 4.4 Results of ML algorithms for lung dataset

ALGORITHM	ACCURACY	PRECISION	RECALL	F1-SCORE
SVM	96.87	97.037	96.43	96.8
Decision Tree	95.83	96	96	95.83
Random Forest	98.95	99.09	98.81	98.94
XGBoost	97.91	98.22	97.62	97.87

As shown in Table 4.4, he SVM algorithm demonstrates robust performance achieving an accuracy rate of 96.87%, a precision of 97.037%, a recall of 96.43%, and an F1-Score of 96.8%, indicating its efficacy in accurately classifying instances and capturing true positives. The Decision Tree model follows suit with a slightly lower accuracy of 95.83% but maintains balanced precision and recall at 96%. Random Forest stands out with exceptional accuracy at 98.95%, coupled with high precision (99.09%) and recall (98.81%), showcasing its prowess in minimizing both false positives and false negatives. XGBoost also performs admirably attaining an accuracy level of 96.87%, precision of 97.037%, recall of 96.43%, and an F1-Score of 96.8%. Overall, these ML algorithms demonstrate strong capabilities in lung cancer classification, with Random Forest exhibiting particularly notable performance across multiple metrics. Figure 4.7 show confusion matrices for lung cancer classification.

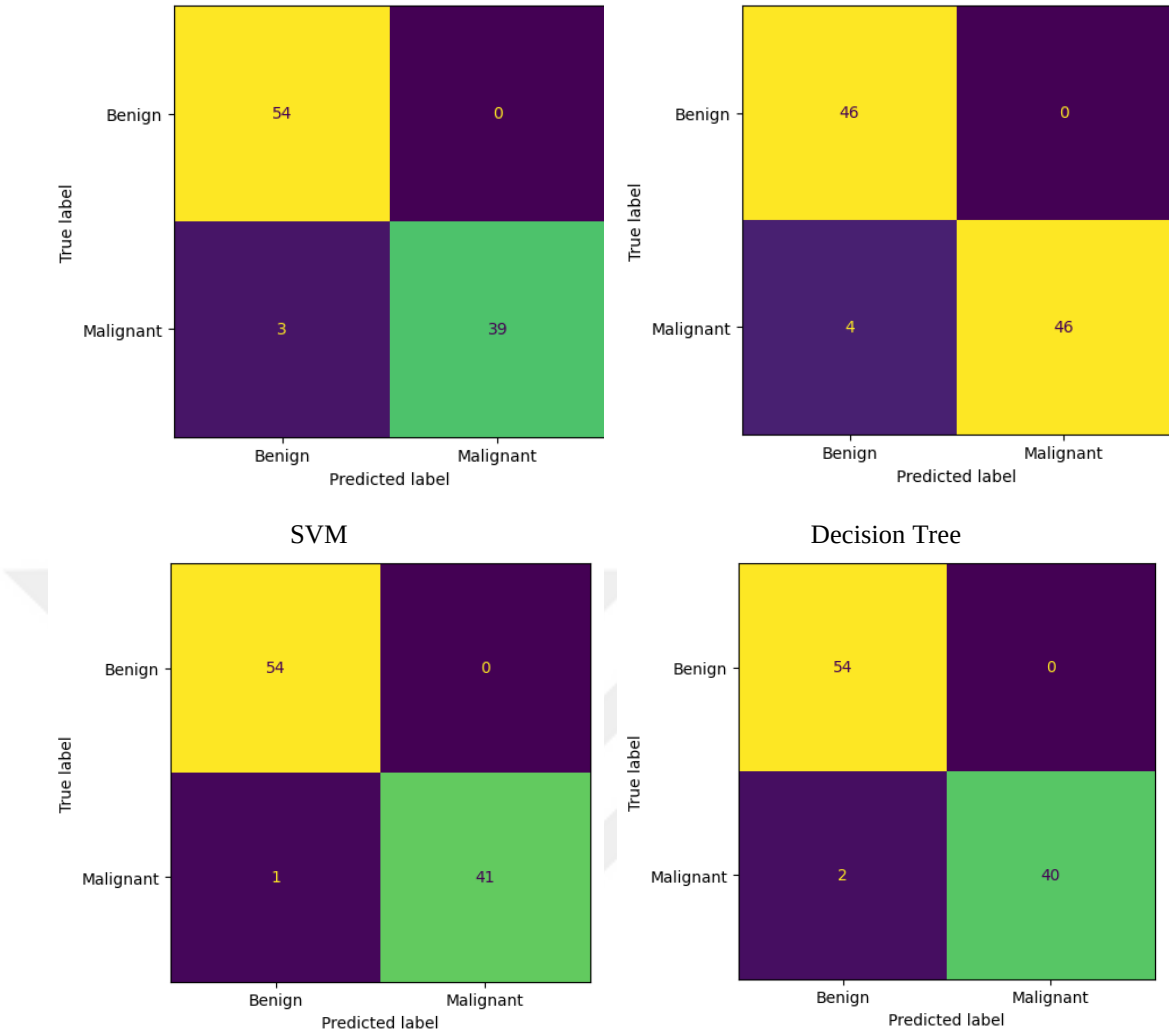


Figure 4.7 Confusion matrices for ML algorithms for lung cancer dataset

As in Figure 4.7, for the SVM model, it correctly classified 54 benign cases and 39 malignant cases, with no false positives or false negatives, indicating a highly accurate and precise classification. The Decision Tree model achieved a balanced performance, correctly classifying 46 benign cases and 46 malignant cases, with no false positives. The Random Forest model demonstrated an exceptional ability to minimize false positives and false negatives, correctly classifying 54 benign cases and 41 malignant cases, with only one misclassification. Similarly, the XGBoost model showcased strong performance, accurately classifying 54 benign cases and 40 malignant cases, with only two misclassifications. These confusion matrices underscore the models' effectiveness in distinguishing between benign and malignant lung cancer cases, with variations in their

error types and an overall trend of limited misclassifications. Figure 4.8 represent The ROC curves and corresponding AUC scores for lung cancer classification showcase the discriminative performance of SVM, Decision Tree, Random Forest, and XGBoost models.

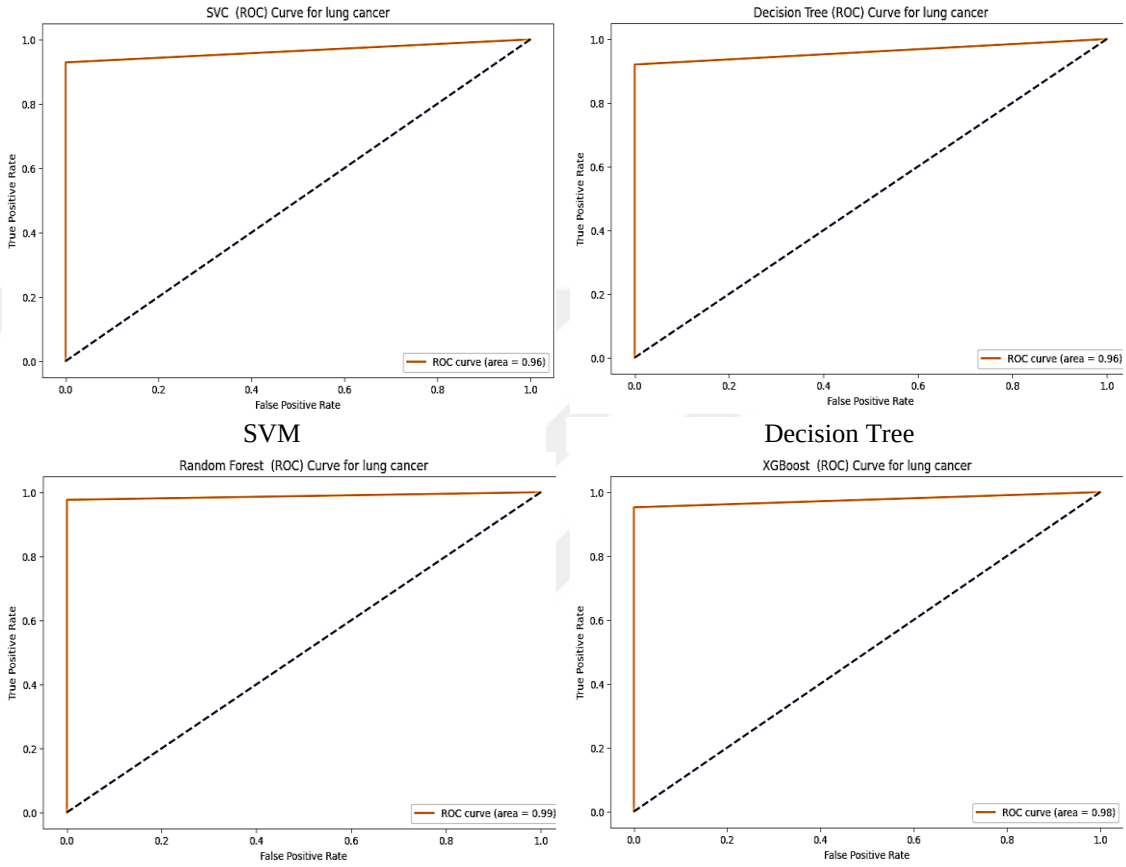


Figure 4.8 ROC AUC scores for the ML algorithms using the lung cancer dataset

Figure 4.8 show that the SVM model demonstrates strong discriminatory ability with an ROC AUC area of 0.96. Similarly, the DT model exhibits an ROC AUC area of 0.96, highlighting its robust ability to distinguish between benign and malignant cases. The RF model stands out with the highest AUC area of 0.99, underlining its exceptional discriminatory power. The XGBoost model also performs admirably with an AUC area of 0.98, further confirming the models' effectiveness in distinguishing between different classes in lung cancer classification.

4.3 Results of Prostate Cancer Dataset

In this part, we represent the results obtained from using both statistical and machine learning algorithms for the three cancer datasets: breast cancer, lung cancer, and prostate cancer.

4.3.1 Results of statistical learning algorithms

In this section, we present the results of applying statistical learning algorithms to the prostate cancer dataset. Table 4.5 provides a comprehensive overview of the outcomes for Lasso logistic regression, Ridge logistic regression, and Elastic Net logistic regression. The table includes essential performance metrics.

Table 4.5 Results of statistical learning algorithms for prostate dataset

ALGORITHM	ACCURACY	PRECISION	RECALL	F1-SCORE
Lasso logistic regression	77.64	79.14	78.78	77.64
Ridge logistic regression	75.29	76.4	76.73	75.28
Elastic Net logistic regression	74.11	74.92	75.34	74.09

As seen in Table 4.5, in terms of Accuracy, Lasso logistic regression achieved the highest at 77.64%, followed by Ridge logistic regression at 75.29%, and Elastic Net logistic regression at 74.11%. Precision is highest for Lasso at 79.14%, followed by Ridge at 76.4% and Elastic Net at 74.92%. Similarly, Recall, which measures the ability to capture all actual positives, is highest for Lasso at 78.78%, followed by Ridge at 76.73%, and Elastic Net at 75.34%. The F1-Score achieving the highest at 77.64%, followed by Ridge at 75.28%, and Elastic Net at 74.09%. Figure 4.9 shows the confusion matrices for prostate cancer classification using Lasso logistic regression, Ridge logistic regression, and ElasticNet logistic regression provide a detailed understanding of their performance on the high and low-volume classes.

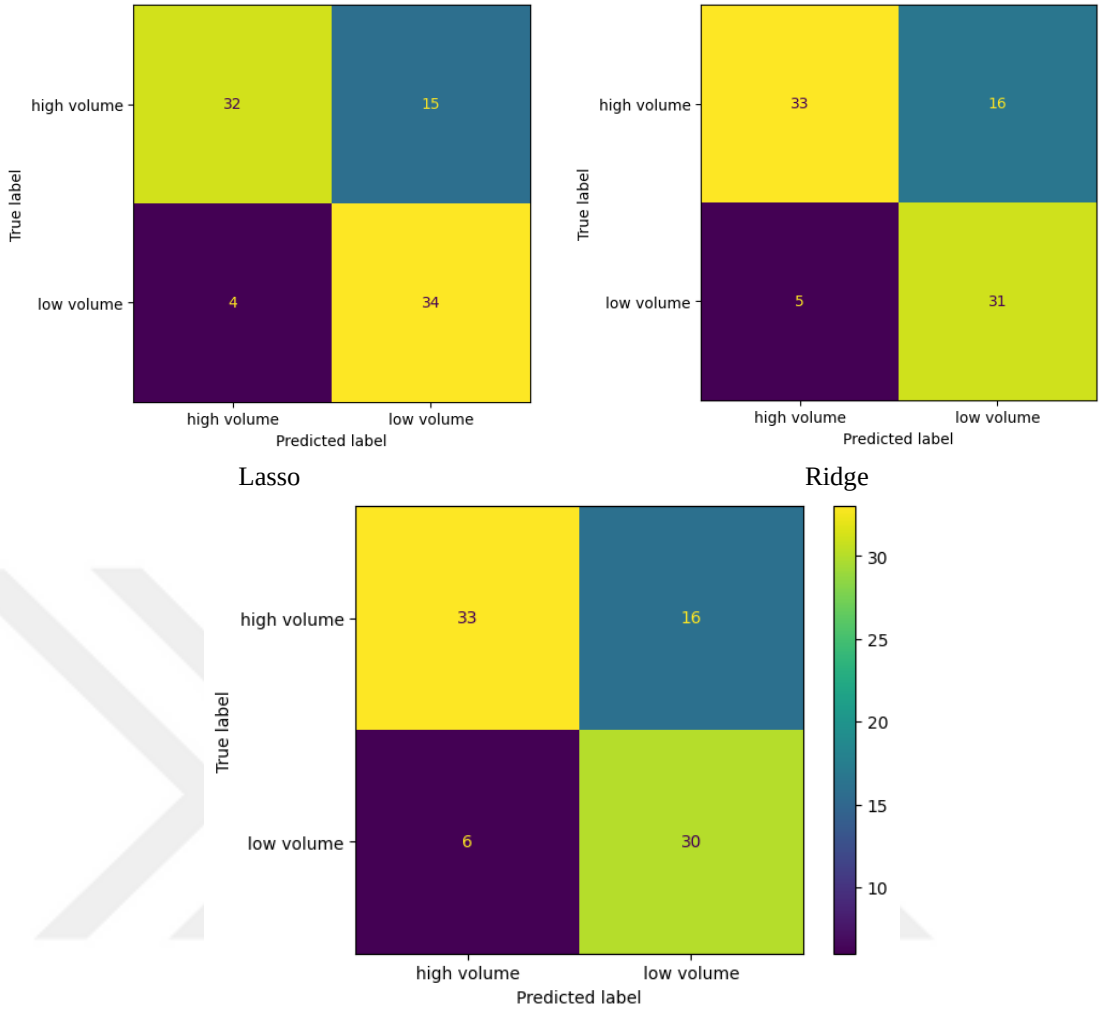


Figure 4.9 Confusion matrices for statistician learning algorithms for prostate dataset

As depicted in Figure 4.9, for the Lasso model, it correctly classified 32 high-volume cases and 34 low-volume cases, with 15 false positives and 4 false negatives. The Ridge model exhibited similar performance, accurately classifying 33 high-volume cases and 31 low-volume cases, with 16 false positives and 5 false negatives. Similarly, the ElasticNet model showcased consistent accuracy, correctly classifying 33 high-volume cases and 30 low-volume cases, with 16 false positives and 6 false negatives. These confusion matrices highlight the models' ability to distinguish between high and low-volume instances, with variations in their error types and an overall trend of balanced performance in prostate cancer classification. Figure 4.10 show The ROC curves and corresponding AUC scores for prostate cancer classification reveal the discriminative

performance of Lasso logistic regression, Ridge logistic regression, and ElasticNet logistic regression.

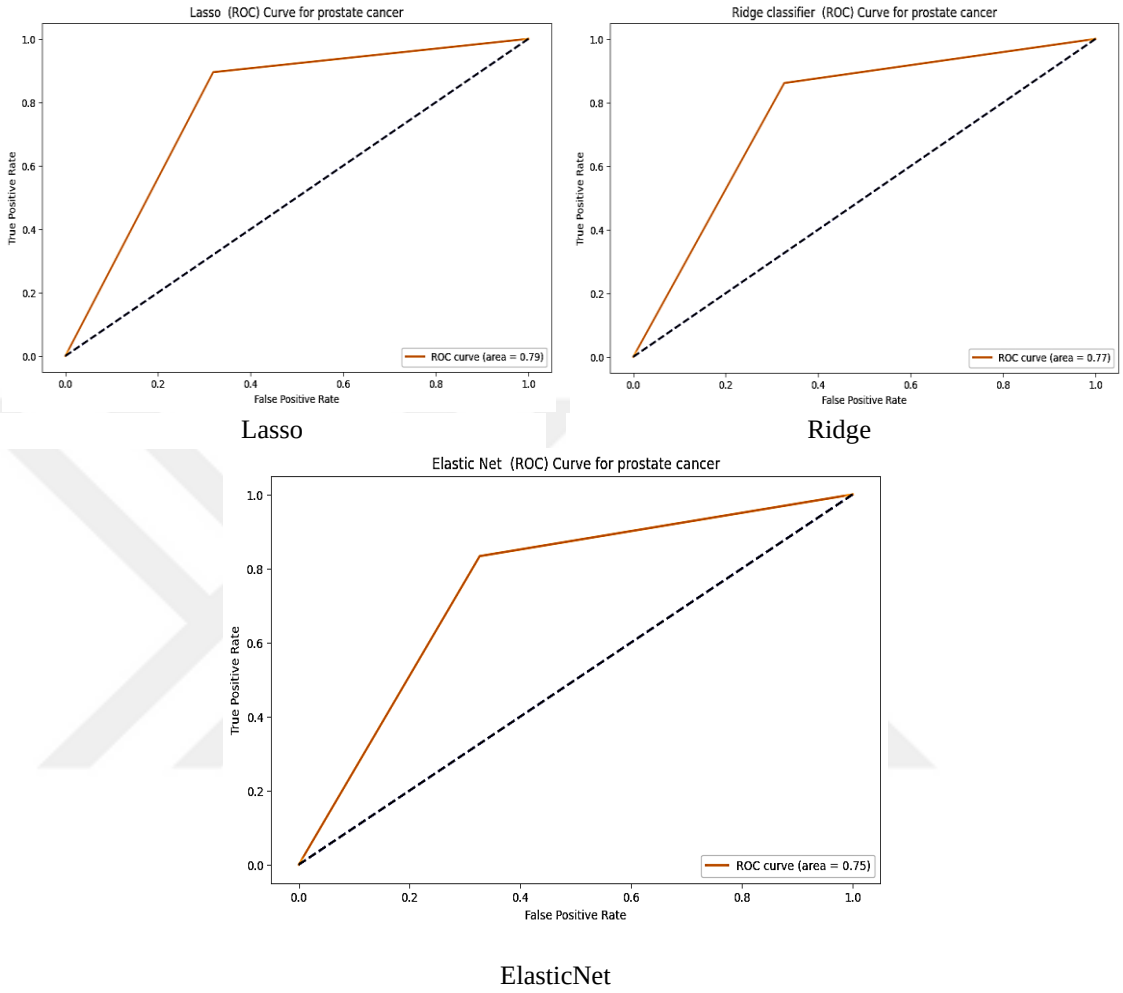


Figure 4.10 ROC AUC scores for the statistical algorithms using the prostate cancer dataset

According to Figure 4.10, the lasso logistic regression achieved an ROC AUC area of 0.79, indicating a moderate ability to distinguish between high and low-volume instances. The Ridge model demonstrated a slightly lower ROC AUC area of 0.77, while the ElasticNet model exhibited an AUC area of 0.75. These AUC scores suggest that all three models have a fair discriminatory power, with Lasso logistic regression performing slightly better than Ridge and ElasticNet in distinguishing between different classes in the prostate cancer dataset. The AUC scores, in conjunction with other performance metrics, provide valuable insights into the models' effectiveness in classifying high and low-volume cases in prostate cancer.

4.3.2 Results of ML algorithms

This section showcases the outcomes obtained by applying diverse ML algorithms to the prostate cancer dataset. The table provides performance metrics. Table 4.6 provides a comprehensive overview of the performance metrics for four machine learning algorithms applied to a prostate dataset: SVM, Decision Tree, Random Forest, and XGBoost.

Table 4.6 Results of ML algorithms for prostate dataset

ALGORITHM	ACCURACY	PRECISION	RECALL	F1-SCORE
SVM	75.29	74.63	75.15	74.79
Decision Tree	63.53	63.75	63.59	63.45
Random Forest	69.41	70	70.57	69.41
XGBoost	67.06	67.22	67.11	67

In terms of Accuracy, SVM leads with 75.29%, followed by Random Forest at 69.41%, XGBoost at 67.06%, and Decision Tree at 63.53%. Precision, which measures the accuracy of positive predictions, is highest for SVM at 74.63%, followed by Random Forest at 70%, XGBoost at 67.22%, and Decision Tree at 63.75%. Recall, representing the ability to capture all actual positives, is led by SVM at 75.15%, followed by Random Forest at 70.57%, XGBoost at 67.11%, and Decision Tree at 63.59%. The F1-Score reflects a similar trend with SVM achieving the highest at 74.79%, followed by Random Forest at 69.41%, XGBoost at 67%, and Decision Tree at 63.45%.

These results indicate varying degrees of performance among the machine learning algorithms, with SVM demonstrating the highest accuracy and F1-Score. Random Forest also shows strong performance across multiple metrics, while Decision Tree lags slightly behind. Figure 4.11 show the confusion matrices for prostate cancer classification using SVM, Decision Tree, Random Forest, and XGBoost models provide a detailed picture of their performance on high and low-volume instances.

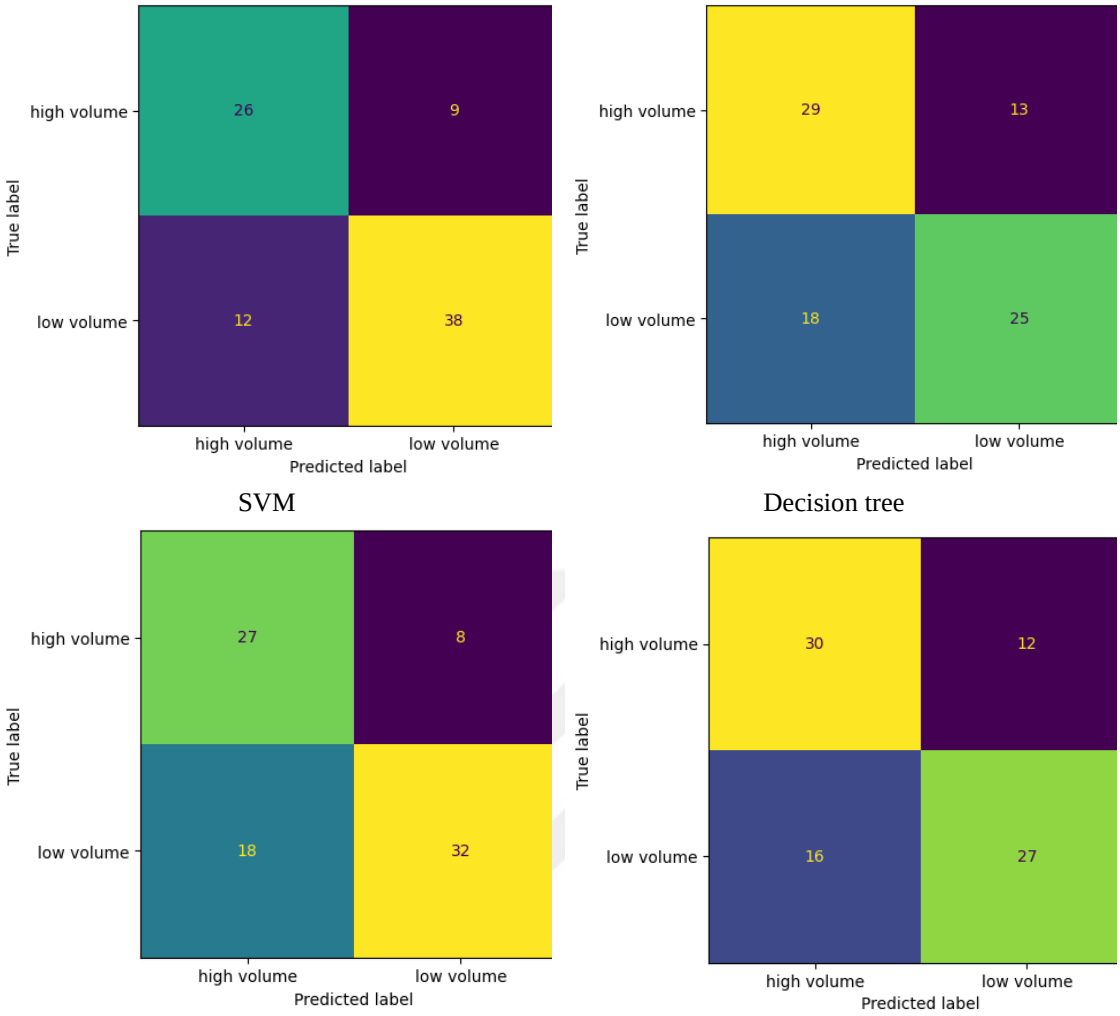
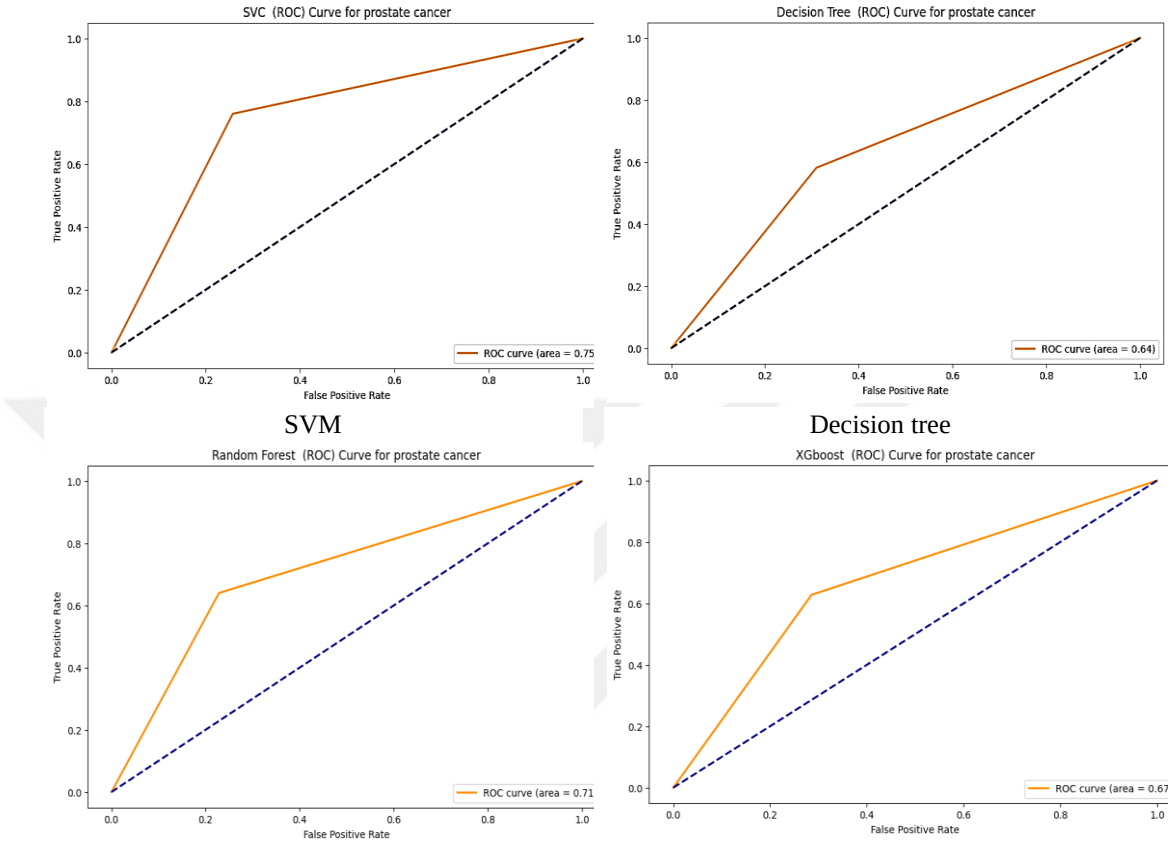


Figure 4.11 Confusion matrices for ML algorithms for prostate cancer dataset

In Figure 4.11, for SVM, it correctly classified 26 high-volume cases and 38 low-volume cases, with 9 false positives and 12 false negatives. The Decision Tree model accurately classified 29 high-volume cases and 25 low-volume cases, with 13 false positives and 18 false negatives. Similarly, the Random Forest model demonstrated good accuracy, correctly classifying 27 high-volume cases and 32 low-volume cases, with 8 false positives and 18 false negatives. The XGBoost model showcased robust performance, accurately classifying 30 high-volume cases and 27 low-volume cases, with 12 false positives and 16 false negatives. These confusion matrices highlight the models' ability to distinguish between high and low-volume instances, with variations in their error types and an overall trend of balanced performance in prostate cancer classification. Figure 4.12

represent the AUC scores for prostate cancer classification demonstrate the discriminative capabilities of SVM, Decision Tree, Random Forest, and XGBoost models.



SVM Decision tree

Random forest XGBoost

Figure 4.12 ROC AUC scores for the ML algorithms using the prostate cancer dataset

According to Figure 4.12, SVM leads with an ROC AUC area of 0.75, indicating a moderate ability to distinguish between high and low-volume instances. Decision Tree exhibits a lower ROC AUC area of 0.64, while Random Forest and XGBoost perform slightly better with AUC areas of 0.71 and 0.67, respectively. These AUC scores suggest that SVM has the strongest discriminatory power among the models, followed by RF, XGBoost, and DT. The AUC scores, along with other performance metrics, provide insights into the models' effectiveness in distinguishing between different classes in the prostate cancer dataset, with SVM demonstrating the highest discriminatory performance.

Both (Shakeel et al. 2020) and the proposed thesis focus on utilizing advanced computing techniques to enhance cancer diagnosis. Shakeel et al.'s research specifically addresses

challenges in feature sets related to lung biomedical data, employing neural computing and soft computing techniques like discrete AdaBoost optimized ensemble learning generalized neural networks. This approach results in high predictive accuracy and a minimal error rate. Similarly, my thesis tackles cancer diagnosis across breast, lung, and prostate cancer datasets through a comprehensive preprocessing approach involving data cleaning, scaling, oversampling, and cross-validation. The assessment of statistical learning and machine learning algorithms highlights their effectiveness in distinguishing between benign and malignant samples across different cancer types. Both studies underscore the crucial role of accurate cancer diagnosis in influencing patient outcomes and guiding treatment strategies, sharing a common objective of advancing cancer detection and diagnosis using sophisticated computational techniques.

5. CONCLUSIONS AND RECOMMENDATION

In this section, we present the conclusions drawn from a comprehensive evaluation and comparison of statistical learning algorithms and ML algorithms on breast cancer, lung cancer, and prostate cancer datasets. The study focused on assessing the algorithms' performance in distinguishing between benign and malignant samples across different cancer types. The results highlight the potential of both statistical learning and ML algorithms for cancer classification, providing valuable insights into their strengths and limitations. Moreover, we offer several recommendations and propose future directions to enhance the accuracy of cancer diagnosis and treatment using advanced machine learning techniques. The implications of this research are significant as it guides researchers and medical professionals in selecting suitable algorithms for specific cancer datasets, ultimately leading to improved cancer detection and patient care.

5.1 Conclusion

This investigation performed an extensive comparison between statistical learning algorithms and machine learning algorithms using datasets for breast cancer, lung cancer, and prostate cancer. The findings demonstrate that both types of algorithms exhibited promising performance in their respective datasets. For breast cancer, statistical learning algorithms, particularly Lasso logistic regression, achieved impressive accuracy, precision, recall, and F1-score, all surpassing 97.2% and this fulfil the answer for RQ2. Similarly, ML algorithms, such as XGBoost achieved competitive results with an accuracy of 98.6% and this fulfil the answer for RQ3 and also the RQ1.

In the context of lung cancer, both statistical learning algorithms and ML algorithms demonstrated good performance, with accuracies ranging from 93 % to 98.6%. Notably, for the ML algorithm, Random Forest outperformed others, achieving an accuracy of 98.95% and this also fulfil the answer for RQ1.

For prostate cancer, both types of algorithms achieved moderate accuracy, ranging from 63.53% to 77.64% but in this dataset the statistical algorithms outperforms ML algorithms particularly the Lasso achieve a high performance which equal to 77.64 with comparison to the best score using ML algorithm which equal to 75.29% using SVM.

At the end of this thesis, it is evident that no single algorithm reigns supreme for diagnosing cancer types. Instead, a set of effective algorithms from ML and statistical learning have been identified to address the diagnostic process. Finally, we can conclude that the ML algorithm are very effective than statistical learning algorithms according to the breast and lung cancers dataset and this fulfil the answer for RQ1 but the statistical algorithms are effective than ML for the prostate cancer and this full fill the answer of RQ4.

5.2 Recommendations and Future Work

Based on these promising results, several recommendations and potential areas for future work have been identified:

- Ensemble Methods: Exploring stacking ensemble methods like bagging and boosting ML algorithms.
- Feature Selection: Extending feature selection techniques will help identify the most informative features for cancer classification, leading to enhanced model performance and reduced overfitting risk.
- Exploring other datasets for the other types of cancer.
- Implementation of other statistical learning algorithms and advanced types of them.

In conclusion, this study highlights the potential of both statistical learning algorithms and ML algorithms for cancer classification. Understanding their strengths and limitations will assist researchers and medical professionals in selecting appropriate algorithms for specific cancer datasets. As the field of machine learning continues to

advance, further research and collaboration with the medical community will contribute to more accurate and reliable cancer diagnosis and treatment.

REFERENCES

Abinash, M. J. and Vasudevan, V. 2019. A hybrid forward selection based LASSO technique for liver cancer classification. In: Lecture Notes in Electrical Engineering (Vol. 511), Springer Singapore, 978–981.

Abunasser, B. S., Al-Hiealy, M. R. J., Zaqout, I. S. and Abu-Naser, S. S. 2022. Breast Cancer Detection and Classification using Deep Learning Xception Algorithm. International Journal of Advanced Computer Science and Applications, 13(7): 223–228.

Ahsan, M. M., Mahmud, M. A. P., Saha, P. K., Gupta, K. D. and Siddique, Z. 2021. Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. Technologies, 9(3).

Akai, H., Yasaka, K., Kunimatsu, A., Nojima, M., Kokudo, T., Kokudo, N., Hasegawa, K., Abe, O., Ohtomo, K. and Kiryu, S. 2018. Predicting prognosis of resected hepatocellular carcinoma by radiomics analysis with random survival forest. Diagnostic and Interventional Imaging, 99(10): 643–651.

Athey, S., Tibshirani, J. and Wager, S. 2019. Generalized random forests. The Annals of Statistics, 47(2): 1148–1178.

BHANDARI, A. 2020. Website: https://medium.com/analytics-vidhya/auc-roc-curve-in-machine-learning-clearly-explained-1849b3fa4bfc. Date of accses: 15.06.2020.

Bousquet, O., Boucheron, S. and Lugosi, G. 2004. Introduction to statistical learning theory. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 3176: 169–207.

Chen, T., He, T. and Benesty, M. 2018. XGBoost : eXtreme gradient boosting. R Package Version 0.71(2): 1–4.

Ferlay, J., Lam, F., Colombet, M., Mery, L., Pineros, M. and Znaor, A. 2020. Global cancer observatory: cancer today. International Agency for Research on Cancer, 1(8): 563-570.

Ganggayah, M. D., Taib, N. A., Har, Y. C., Lio, P. and Dhillon, S. K. 2019. Predicting factors for survival of breast cancer patients using machine learning techniques. BMC Medical Informatics and Decision Making, 19(1): 1–17.

Geetha, R., Sivasubramanian, S., Kaliappan, M., Vimal, S. and Annamalai, S. 2019. Cervical cancer identification with synthetic minority oversampling technique and pca analysis using random forest classifier. Journal of Medical Systems, 43(9).

Gupta, P., Chiang, S. and Sahoo, P. K. 2019. Prediction of colon cancer stages and survival. Cancers, 1–16.

Hastie, T., Tibshirani, R. and Wainwright, M. 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1).

Hoerl, A. E. and Kennard, R. W. 2000. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 42(1): 80–86.

Hong, Z. Q. and Yang, J. Y. 1991. Optimal discriminant plane for a small number of samples and design method of classifier on the plane. Pattern Recognition, 24(4): 317–324.

Hossin, M. and bin Sulaiman, M. N. 2015. A Review On Evaluation Metrics For Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process, 5: 1–11.

Wolberg, W., William H., Mangasarian, D., Olvi, S., Nick N. and Street,W. 1995. Breast cancer wisconsin (diagnostic) dataset. UCI Machine Learning Repository.

KEGGLE. 2018. Website: https://www.kaggle.com/datasets/sajidsaifi/prostate-cancer. Date of Access: 06.02.2018

Lyu, L., Yu, H., Ma, X., Chen, C., Sun, L., Zhao, J., Yang, Q. and Yu, P. S. 2022. Privacy and robustness in federated learning: Attacks and defenses. IEEE Transactions on Neural Networks and Learning Systems. 1-20.

Moncada-Torres, A., van Maaren, M. C., Hendriks, M. P., Siesling, S. and Geleijnse, G. 2021. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. Scientific Reports, 11(1): 1–13.

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A. and Brown, S. D. 2004. An introduction to decision tree modeling. Journal of Chemometrics, 18(6): 275–285.

Nageswaran, S., Arunkumar, G., Bisht, A. K., Mewada, S., Kumar, J. N. V. R. S., Jawarneh, M. and Asenso, E. 2022. Lung cancer classification and prediction using machine learning and image processing. BioMed Research International. 1–7

National Cancer Institute. 2022. Website: https://www.cancer.gov/about-cancer/causes-prevention/genetics. Date of access: 17.08 2022.

Noble, W. S. 2006. What is a support vector machine? Nature Biotechnology, 24(12): 1565–1567.

Picard, R. R. and Berk, K. N. 1990. Data splitting. American Statistician, 44(2): 140–147.

Rinesh, S., Maheswari, K., Arthi, B., Sherubha, P., Vijay, A., Sridhar, S., Rajendran, T. a Waji, Y. A. 2022. Investigations on brain tumor classification using hybrid machine learning algorithms. Journal of Healthcare Engineering, 2022: 1–9.

Rohan, T. I., Awan-Ur-Rahman, Siddik, A. B., Islam, M. and Yusuf, M. S. U. 2019. A precise breast cancer detection approach using ensemble of random forest with adaboost, In: 5th International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering, 13(13): 10–14.

OJALA and GARRIGA. 2011. Website: https://scikit-learn.org/stable/modules/cross_validation.html, Date of Access: 12.05.2010.

Shafi, A. S. M., Molla, M. M. I., Jui, J. J. and Rahman, M. M. 2020. Detection of colon cancer based on microarray dataset using machine learning as a feature selection and classification techniques. SN Applied Sciences, 2(7): 1–8.

Shakeel, P. M., Tolba, A., Al-Makhadmeh, Z. and Jaber, M. M. 2020. Automatic detection of lung cancer from biomedical data set using discrete AdaBoost optimized ensemble learning generalized neural networks. Neural Computing and Applications, 32(3): 777–790.

Sharifmoghadam, M. and Jazayeriy, H. 2019. Breast Cancer Classification Using AdaBoost- Extreme Learning Machine. 5th Iranian Conference on Signal Processing and Intelligent Systems, ICSPIS 2019: 1–5.

STEYN, P. 2022. Website: https://towardsdatascience.com/why-is-statistics-important-in-data-science-machine-learning-and-analytics-92b4a410f686. Date of access: 17.03.2022 .

Tewari, Y., Ujjwal, E. and Kumar, L. 2022. Breast cancer classification using machine learning. 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering, 2022: 168–171.

Tran, K. A., Kondrashova, O., Bradley, A., Williams, E. D., Pearson, J. V. and Waddell, N. 2021. Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Medicine, 13(1).

WHO. 2019. Website: https://www.who.int/activities/promoting-cancer-early-diagnosis. Date of access: 30.09.2019

Yifan, D., Jialin, L. and Boxi, F. 2021. Forecast model of breast cancer diagnosis based on RF-adaboost, In: 2021 IEEE 3rd International Conference on Communications, Information System and Computer Engineering, 716–719.

Zou, H. and Hastie, T. 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. Series B: Statistical Methodology, 67(2): 301–320.

CURRICULUM VITAE

Personal Information

Name and Surname : Asmaa Salim Hussaien ALWAZY

Education

MSc	Çankırı Karatekin University Graduate School of Natural and Applied Sciences Department of Electronics and Computer Engineering	2021-2024
Undergraduate	University of Mosul Faculty of Engineering Department of Electronics and Computer Engineering	2010-2014