ÇUKUROVA UNIVERSITY

INSTITUTE OF NATURAL AND APPLIED SCIENCES

M.Sc. THESIS

Analysis of leukemia Cancer Classification with Supervised Machine Learning and Deep Reinforcement Learning Based on gene expression monitoring (via DNA microarray)

Zaid Mohammed Ibrahim IBRAHIM

Department of Electrical and Electronics Engineering

January, 2023

ÇUKUROVA UNIVERSITY

INSTITUTE OF NATURAL AND APPLIED SCIENCES

M.Sc. THESIS

Analysis of leukemia Cancer Classification with Supervised Machine Learning and Deep Reinforcement Learning Based on gene expression monitoring (via DNA microarray)

Zaid Mohammed Ibrahim IBRAHIM

Department Of Electrical And Electronics Engineering

This Master's Thesis was evaluated by the following Jury Members on 23/01/2023 and was approved by unanimity / majority of votes.

Jury : Prof. Dr. Ulus ÇEVİK (Supervisor)

: Prof. Dr. Turgay İBRİKÇİ (Co- Supervisor)

: Assoc. Dr. Sami ARICA

: Assoc. Dr. Kerem ÜN

: Assist. Prof. Erkut TEKELİ

This Thesis was written in the Department of Electrical and Electronics Engineering of Institute of Natural and Applied Sciences.
Thesis Number:

Prof. Dr. Sadık DİNÇER
Director
Institute of Natural and Applied Sciences

CONTENTS

CUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

M.Sc. THESIS

Analysis of leukemia Cancer Classification with Supervised Machine Learning and Deep Reinforcement Learning Based on gene expression monitoring (via DNA microarray)

Zaid Mohammed Ibrahim IBRAHIM

Department of Electrical and Electronics Engineering

Supervisor: Prof. Dr. Ulus ÇEVİK
Co- Supervisor: Prof. Dr. Turgay İBRİKÇİ

ABSTRACT

Cancer is a progressively common risk that endangers human health around the globe, and an early diagnosis necessitates a positive prognosis. A substantial quota of the world population dies because of cancer, and the research community is alarmed that by 2025 a probable increase of cancer cases will reach a staggering 25 million cases. Several studies have already proven cancer genesis is mostly supported by an accretion of dangerous mutations, amplifying the recognition of cancer based on genome information. Moreover, according to health practitioners, cancer has been stated as one of the most lethal illnesses of the genome. It has been the pivot of research by doctors, pathologists, biologists, data scientists, and other life science and health professionals as it is a need of the hour to have an advanced automated method that can quickly identify cancer and its subtype, i.e., Acute Lymphocytic Leukemia (ALL) and Acute Myelocytic Leukemia (AML). This research aims to identify leukemia cancer subclasses using DNA Gene Expression. Much research can be found using gene expression data in the cancer classification domain. All the research can be categorized into four major categories: traditional data mining methods, Deep Learning (DL) methods, ML methods, and Classification using Deep Reinforcement Learning (RL); the most famous datasets are The Cancer Genome Altus and The Gene Expression Dataset. However, most of the literature utilized one type of model or some combination of different models. To the best of our knowledge, the optimal compassion of Machine Learning (ML) models, DL models, and Deep-RL models are still unknown or unavailable in one place. This thesis aims to explore and implement the models based on ML, DL, and Deep-RL and provide a comparative analysis. A comparative analysis of the three feature selection techniques was performed to show the importance of feature selection in Leukemia cancer prediction. Based on the analysis result, Logistic Regression (LR) gave the highest accuracy of 97% with the raw dataset, while with PCA version of the dataset Support Vector Machine (SVM), Gaussian-NB, and LR obtained an accuracy of 91.%. Furthermore, with the Random Forest (RF) importance dataset version, SVM achieved higher accuracy of 97% along with lasso regularization dataset version of DNN performed well and obtained a higher accuracy of 97%. Out of six different Deep-RL models with PCA dataset, model-3,5,6 achieved better accuracy of 88.24%. Although the PCA version of six Deep-RL configurations, model-1 reached an AUROC value of 72, which is higher than any non-PCA (raw dataset) version of six Deep-RL models. Moreover, with RF importance dataset out of six different configurations of Deep-RL model 6 obtained higher accuracy of 88.24%, and with lasso regularization dataset version six different configurations of Deep-RL model 1 achieved higher accuracy of 73.53%

Keywords: Deep Reinforcement Learning, Deep learning, Machine Learning, Gene Expression; Leukemia Cancer Classification; PCA; Random Forest Importance, Lasso Regularization

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

Denetimli Makine Öğrenimi ve Gen İfade İzlemeye Dayalı Derin Takviyeli Öğrenme ile Lösemi Kanseri Sınıflandırmasının Analizi (DNA mikrodizisi aracılığıyla)

Zaid Mohammed Ibrahim Ibrahim

Elektrik - Elektronik Mühendisliği Anabilim Dalı

Danışman: Prof. Dr. Ulus ÇEVİK
İkinci Danışman: Prof. Dr. Turgay İBRİKÇİ

ÖZ

Kanser, dünya genelinde insan sağlığını tehlikeye atan, giderek yaygınlaşan bir risktir ve erken teşhis için olumlu bir prognoz (hastalığın sonucunu tahmini) gerektirmektedir. Dünya nüfusunun önemli bir çoğunluğu kanserden ölmektedir ve araştırma topluluğu, 2025 yılına kadar olası bir kanser vakası artışının şaşırtıcı bir şekilde 25 milyon vakaya ulaşacağı konusunda alarma geçmiştir. Birkaç çalışma, kanser oluşumunun çoğunlukla tehlikeli mutasyonların artmasıyla desteklendiğini kanıtlamıştır ve bu da genom bilgisine dayalı kanserin tanınmasını güçlendirmektedir. Ayrıca, sağlık pratisyenlerine göre kanser, genomun ölümcül hastalıklarından biri olarak belirtilmiştir. Doktorlar, patologlar, biyologlar, veri bilimcileri ve diğer yaşam bilimleri ve sağlık profesyonelleri tarafından, kanseri ve alt tipini, yani Akut Lenfositik Lösemi ve Akut Miyelositik Lösemi hastalıklarını hızlı bir şekilde tanımlayabilen gelişmiş bir otomatik yönteme sahip olmak saatlik bir ihtiyaç olması nedeniyle araştırmaların merkezi olmuştur. Bu araştırma, DNA Gen Ekspresyonunu kullanarak lösemi kanseri alt sınıflarını belirlemeyi amaçlamaktadır. Gen ifadeleri verileri kullanılarak kanser sınıflandırma alanında birçok araştırma bulunabilir. Tüm araştırmalar dört ana kategoriye ayrılabilir: geleneksel veri madenciliği yöntemleri, derin öğrenme yöntemleri, ML yöntemleri ve Derin Takviyeli Öğrenme kullanılarak Sınıflandırma; en ünlü veri kümeleri The Cancer Genome Altus ve The Gen Expression Dataset'tir. Bununla birlikte, literatürün çoğu, bir tür model veya farklı modellerin bir kombinasyonunu kullanmıştır. Bildiğimiz kadarıyla, Makine Öğrenimi modellerinin, Derin Öğrenme modellerinin ve Derin Takviyeli Öğrenme modellerinin optimal anlayışı hala bilinmemektedir veya tek bir yerde mevcut değildir.

Bu tez, makine öğrenimi, derin öğrenme ve derin pekiştirmeli öğrenme tabanlı modelleri keşfetmeyi ve uygulamayı ve tüm çalışmalara ve uygulamaya tek bir yerde yardımcı olmak için kapsamlı bir karşılaştırmalı analiz sağlamayı amaçlamaktadır. Lösemi kanseri tahmininde özellik seçiminin önemini göstermek için 3 özellik seçim tekniğinin karşılaştırmalı analizi yapıldı.

SVM, Gaussian-NB ve LR veri kümesinin PCA sürümü %91, doğruluk elde etti. Ayrıca, rasgele forest importance veri seti versiyonu ile SVM, %97 gibi daha yüksek doğruluk elde etti ve DNN'nin kement düzenlileştirme veri seti versiyonu iyi performans gösterdi ve %97 gibi daha yüksek bir doğruluk elde etti. PCA veri setine sahip altı farklı Deep-RL modelinden model-3,5,6, %88,24 oranında daha iyi doğruluk elde etti. Altı Deep-RL konfigürasyonunun PCA versiyonu olmasına rağmen, model-1, altı Deep-RL modelinin herhangi bir PCA olmayan (ham veri seti) versiyonundan daha yüksek olan 72 AUROC değerine ulaştı. Ayrıca, Deep-RL model 6'nın altı farklı konfigürasyonundan forest importance veri seti ile %88,24 gibi daha yüksek doğruluk elde etti ve Deep-RL model 1'in altı farklı konfigürasyonu ile kement düzenlileştirme veri seti versiyonu ile %73,53 daha yüksek doğruluk elde etti.

Anahtar Kelimeler: Derin Takviyeli Öğrenme, Derin öğrenme, Makine öğrenme, Gen Ekspresyonu; Lösemi Kanseri Sınıflandırması; PCA; Rastgele Orman Önemi, Lasso Düzenlileştirme

GENİŞLETİLMİŞ ÖZET

Depolama sunucularına günlük olarak çok fazla veri indiriliyor. Buradan, uygunluğunu hafifletmek için belirli bir açıdan temsil edilen tek göreli verileri rafine etmek için bir yöntem aramanın kaçınılmazlığı ortaya çıkıyor. Genellikle, veri madenciliği, stratejik analiz için veritabanından çıkarılan güncellenmiş verileri elde etmek için kullanılır. Tıp alanında veri madenciliği, muazzam kapasiteye sahip etkili bir teknolojidir ve hastalara, tıbbi bakım topluluklarına, araştırmacılara ve sigortacılara avantajlar sunar.

Makine öğrenimi, açık programlama olmadan önceden keşfedilmemiş içgörüleri ortaya çıkarmak için örüntü tanıma tekniklerini kullanma yeteneğine sahiptir. Bu yetenek onkoloji alanında, özellikle genomik veriler kullanılarak çeşitli kanser türlerinin sınıflandırılmasında uygulanmıştır. Bu yaklaşım, geleneksel biyolojik metodolojilerden bir ayrılmayı temsil eder.

Ayrıca lösemi hücreleri morfolojik görünümlerine göre sınıflandırılmıştır. Yorucu, pahalı ve zaman alan bir prosedür olan tümör hücreleri arasındaki farklılıkları tanımak için son derece yetkin kaynaklar gereklidir. Temel kaynaklara yeterli erişilebilirlik olsa bile, yöntemin güvenli olmadığını söylemek oldukça teşvik edicidir. Geleneksel yöntemin bu sınırları, belki de hücre sınıflandırması için bir temel olarak kullanılan diğer parametrelerin sınıflandırılması gerekliliği ile sonuçlandı. Gen Ekspresyonu verileri, alt sınıflandırma çalışmaları için değerli bilgiler sunabilir. Buna karşılık, DNA mikrodizileri, bir sürü genin gen ekspresyonu verilerinin eş zamanlı olarak gözlemlenmesinde önemli bir rol oynamıştır. Anormal hücrelere kıyasla normal hücrelerde genlerin göreceli ekspresyon seviyelerini yönetmesi muhtemeldir. İfade düzeylerinin incelenmesi, söz konusu hücrelerin gen imzasına göre kategorizasyon kararlarının oluşturulmasında değerli anlayışlarla sonuçlanabilir.

Bu çalışma, farklı modelleri eğitmek için çeşitli veri parametrelerini kullanma yöntemlerini karşılaştırdı. Küçük gen setlerini şematik sınıflandırma düzenlemelerinde kullanmanın dezavantajı, kararın küçük bir veri setine göre verilmesi ve yapının tutarlılığı hakkında sorular sorulmasıdır. Bu arada, gen Ekspresyonu bir sürü gereksinime dayanır; bazıları tanımlanırken diğerleri tanımlanamıyor; Oldukça küçük bir gen ifadesi veri setine tabi bir karara varmak asla ihtiyatlı değildir. Büyük gen Ekspresyonu veri setlerinin kullanımında, başka zorluklara da katlanırlar. Bir karara dayandırmak için yeterli verinin erişilebilir olduğunu garanti etse de, büyük veri kümelerini kullanmanın dezavantajı, makine öğrenimi algoritmalarının artan karmaşıklığı ve sistemdeki yüksek sayısal istikrarsızlık olasılığıdır.

Bu çalışmanın temel amacı, kanser sınıflandırması için bir kanser genomu veri seti kullanarak çoklu sınıflandırıcıların performansını değerlendirmektir. Çalışma, farklı sınıflandırıcıların performansını karşılaştırmanın yanı sıra, boyut azaltma için Rastgele Orman Önemi, Lasso Düzenlileştirme ve PCA gibi tekniklerin kullanımını da incelemektedir.

Çalışmanın sonuçları, denetimli öğrenme algoritmalarının bir kanser genomu veri setine uygulandığında kanser sınıflandırması için uygun olduğunu ve bu tekniklerin kullanılmasının hastalar için gereksiz teşhis testlerinden kaçınmaya yardımcı olabileceğini düşündürmektedir.

ACKNOWLEDGMENTS

I want to express my gratitude to Prof. Dr. Ulus CEVIK for his encouragement and assistance during the academic year. I am thankful to him for his guidance, encouragement, and constructive criticism during my thesis writing.

Also, I would want to express my heartfelt appreciation to Prof. Dr. Turgay İBRİKÇİ. I cannot express my gratitude enough for his unwavering support and assistance. I owe him a debt of gratitude for his supervision, guidance, encouragement, constructive counsel, and unwavering belief in my ability to complete my thesis. Every time I attend one of his sessions, I leave feeling energized and inspired. This thesis would not have come to fruition if it had not been for his support and assistance.

I want to express my heartfelt gratitude and appreciation to the Turkish government, the Presidency for Turks Abroad and Related Communities (YTB), and the Turkish people for providing me with this opportunity and their kindness and hospitality.

Finally, I wish to express my heartfelt thankfulness to my family for their patience, encouragement, and continuous moral support.

LIST OF TABLES

LIST OF FIGURES

LIST ABREVIATIONS

ALL	: Acute Lymphocytic Leukemia
AML	: Acute Myelocytic Leukemia
AI	: Artificial Intelligence
ANN	: Artificial Neural Network
BM	: Bone Marrow
CNN	: Convolution Neural Network
DT	: Decision Tree
DL	: Deep Learning
DNN	: Deep Neural Network
Deep-RL	: Deep Reinforcement Learning
Gaussian-NB	: Gaussian Naive Baye Gaussian Naive Bayes
GDL	: Genome Deep Learning
GB	: Gradient Boosting
GT	: Ground Truth
K-NN	: K-Nearest Neighbor
LR	: Logistic Regression
ML	: Machine Learning
NN	: Neural Network
PCA	: Principal Component Analysis
PML-RARA	: Promyelocytic Leukemia/Retinoic Acid Receptor Alpha
RF	: Random Forest
ROC	: Receiver Operating Curve
RNN	: Recurrent Neural Network
RBC	: Red Blood Cells
RL	: Reinforcement Learning
SVD	: Singular Value Decomposition
SVM	: Support Vector Machine
TCGA	: The Cancer Genome Atlas
WBCs	: White Blood Cells

1. INTRODUCTION

1.1. Preliminary Background

In recent times image processing approaches have been extensively utilized to identify numerous medical diseases. Leukemia is an extremely intriguing research field since it relates to the class of blood cancer, which might affect individuals of all ages and even children (Ratley et al., 2020). The utilization of the image process, along with Computer-based algorithms, makes the classification feasible. When the expert in the field performs leukemia disease detection, errors might be exhibited because of a lack of information or inaccurate information in microscopic images. Therefore, a Computer-Based algorithm might be a valuable area to increase detection accuracy (Facts Spring, 2014). In human blood, there are two sorts of White Blood Cells (WBCs) present, and once the affected cells are monocytes and granulocytes, leukemia should be classified as myelogenous.

According to health practitioners, leukemia is a malignant (cancer) disease that damages bone marrow and blood caused by blood clotting and uncontrolled blood cell accumulation (Facts Spring, 2014). This disease destroys blood-producing cells within the bone marrow. Acute leukemia cancer spreads more rapidly. In such circumstances, the cells of bone marrow do not develop properly. Acute leukemia is classified into two types:

- Acute Myelocytic Leukemia activates within early myeloid cells. That type of cell develops into WBC aside from red blood cells, lymphocytes, or platelet-producing cells. This is the most familiar kind of leukemia among the elderly. For most cases, early detection of this cancer might steer to effective treatment. In this sort of cancer, a person might feel the issue with breathing, bleeding, etc. (Oostindjer et al., 2014).
- Acute Lymphocytic Leukemia (Deb & Raji Reddy, 2003) starts in the lymphocyte, White Blood Cells. The WBC plays an essential role in the immune system. This kind accounts for about 80% of all leukemia cases in childhood.

1.2. Acute Lymphoblastic Leukemia

The ALL affects children primarily. When DNA abnormalities are taken by stem cells of bone marrow, that causes overproduction of cells (Pui, 2006). Consequently, immature stem cells, known as lymphoblasts, overflow the body, causing bleeding from the nose or gums, infections, weakness, enlarged lymph nodes, breath shortness, fever, bruising, and bone pain (O'Brien et al., 2018). Even though ALL is curable, early discovery is essential for the survival of patients. Doctors require several tests to confirm the findings in the laboratory. Some tests can be invasive, requiring a lumbar puncture or biopsy of bone marrow. Other tests can be less invasive, requiring a single peripheral blood sample. The minimally invasive test is known as a complete blood count. A blood

sample is run with the hematology analyzer, which develops quantitative findings during such a test (Narayanan & Shami, 2012). Figure 1.1 depicts the thrombocytes.

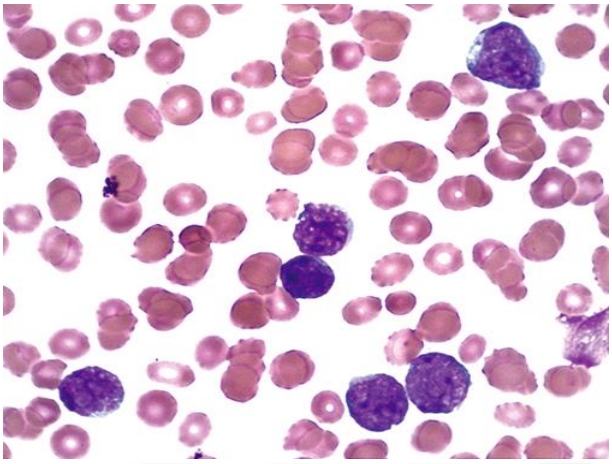

Figure 1.1. The blueish-purple color of ALL (Rahman et al., 2021)

Furthermore, abnormal discoveries need additional inspection using a microscope. During the inspection, a laboratorian examines the different kinds of cells & records qualitative observations such as cell shape. The bloodstain is usually put to a peripheral blood smear to distinguish them under the microscope. The Red Blood Cells (RBC) are more frequent than other parts and seem to be the greyish-pink biconcave disk. The WBCs are classified into five kinds based on the dark blue-purple staining nucleus: lymphocyte, neutrophil, basophil, eosinophil, and monocyte are all types of WBCs. In ALL cases, lymphoblasts may be seen under the microscope. Lymphoblasts also have a distinct morphology from normal lymphocytes. However, they differ in appearance. A typical lymphocyte is about the size of an RBC & contains a spherical nucleus that stains blue and purple. The nucleus is devoid of nucleoli, densely packed with closed chromatin, & has smooth borders. The cytoplasm is typically light blue & sparse, although it might be copious depending on the cell's responsiveness.

On the other hand, a lymphoblast could have a more prominent nucleus that dyes sparsely red purple. A nucleus might have distinct nucleoli, open chromatin, & be indented by rough borders. The cytoplasm could be vivid blue, but it is generally sparse. There is significant variance across ALL the subtypes (Terwilliger & Abdul-Hay, 2017). It is challenging to identify normal lymphocytes and aberrant lymphoblasts consistently and objectively. The precision of the findings is affected by the observer's competence, dedication, and blood sample quality.

1.3. Acute Myelocytic Leukemia

AML can be described as abundant acute leukemia in adults, which worsens swiftly if left untreated. AML is also named acute non-lymphocytic leukemia or acute myelogenous leukemia (Rai et al., 1981). The bone marrow typically produces stem cells of blood known as immature cells,

which mature with time. The blood stem cell can be classified into lymphoid, myeloid, or stem cell (Ablin, 1972). Typically, the lymphoid blood stem cell matures into a white blood cell. In the AML, the Myeloid blood stem cells develop into a myeloblast, which can be explained as an immature WBC. In AML, an odd number of stem cells might transform into aberrant RBCs or platelets (McKenzie, 2005). These defective WBC, platelets, and RBCs are blasts and leukemia cells. Leukemia cells may accumulate in bone marrow & blood, making space for healthy WBC, platelets, and RBC. Infection, anemia, or superficial bleeding are the symptoms.

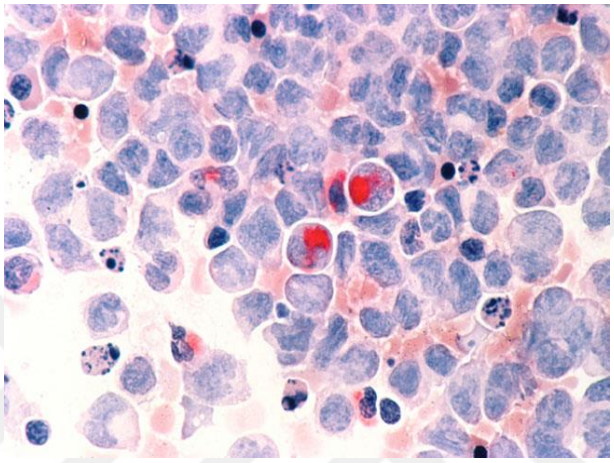

Figure 1.2. The red-purple color of AML (MacCallum, 2022)

Leukemia cells can move outside the bloodstream to other parts of the nervous system like the spinal cord, brain, gums, and skin. Leukemia cells may sometimes produce a solid tumor termed myeloid sarcoma (Miyoshi et al., 2018). Most AML subgroups are determined by how advanced cancer cells become during diagnosis time, depicted in Figure 1.2, and how they vary from the normal cells. Furthermore, Acute promyelocytic leukemia is also a type of AML. This cancer develops when chromosome 15 genes exchange places with chromosome 17 genes, resulting in an irregular gene known as PML-RARA. This PML-RARA gene delivers a signal that inhibits promyelocytes' maturation, which is a kind of WBC. Severe bleeding & blood clots are possible complications (Gorin, 1998).

1.4. Initial Symptoms of Leukemia

Leukemia might present in numerous behaviors. Frequently, the symptoms are unspecific, including fever, weight loss, and poor appetite. Infrequently, some symptoms are affiliated with Bone Marrow (BM) deficiency, for instance, pallor and bruising. Figure 1.3 illustrates the usual acute and chronic leukemia symptoms that could affect the skin, lungs, muscles, joints, bones, and bones (Huang et al., 2014). These particular symptoms are generic for leukemia and frequently, when the full constellation is not present, might be inaccurate for other benign disorders, for instance, viral infection.

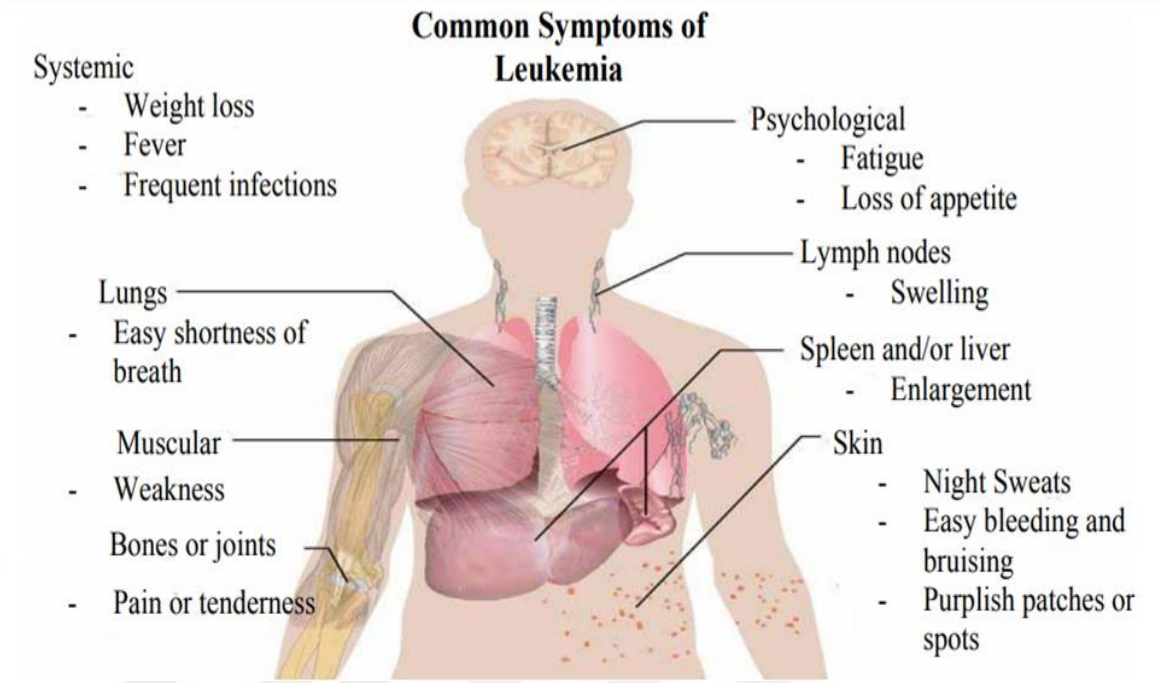

Figure 1.3. Common Symptoms of Leukemia (MedicalStay Group, 2014)

1.5. Diagnosis of Acute Leukemia

Diagnosing acute leukemia needs numerous laboratory tests. The doctor goes through the patient's medical history to confirm how long the symptoms have been present. After that, the patient has to perform a routine physical examination to check abnormalities such as bleeding and enlarged lymph nodes. A microscopic morphological examination will be requested if the doctor doubts acute leukaemia. Based upon the microscopic morphological examination results, additional laboratory tests such as BM aspirate morphological examination, cytogenetics analysis, and immunophenotyping would be required (*Childhood Leukemia Early Detection, Diagnosis, and Types*, 2022). Figure 1.4 and 1.5 depicts the crucial steps that are obligatory to be taken by a doctor to diagnose a patient with acute leukaemia.

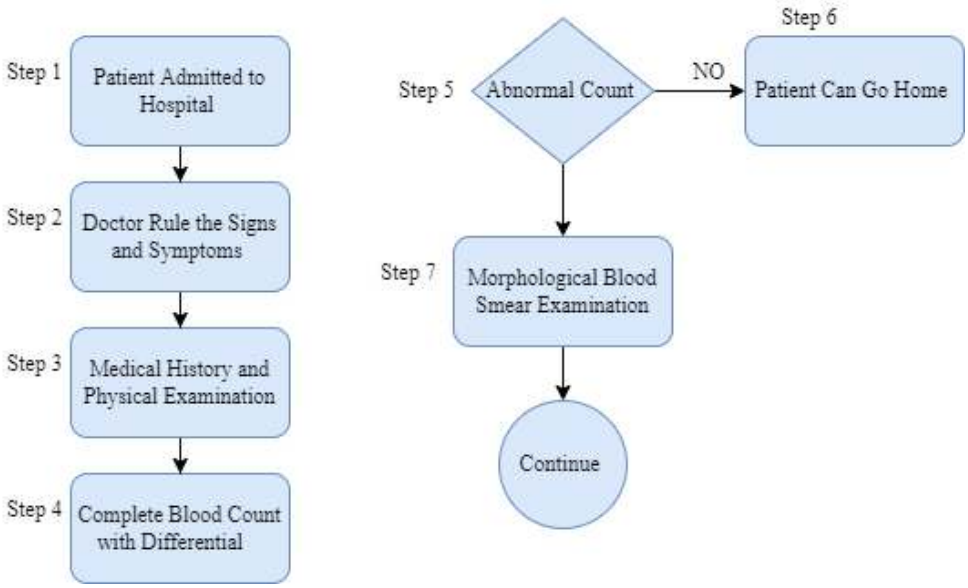

Figure 1.4. Steps to Confirm Acute Leukemia Diagnosis (Part 1)

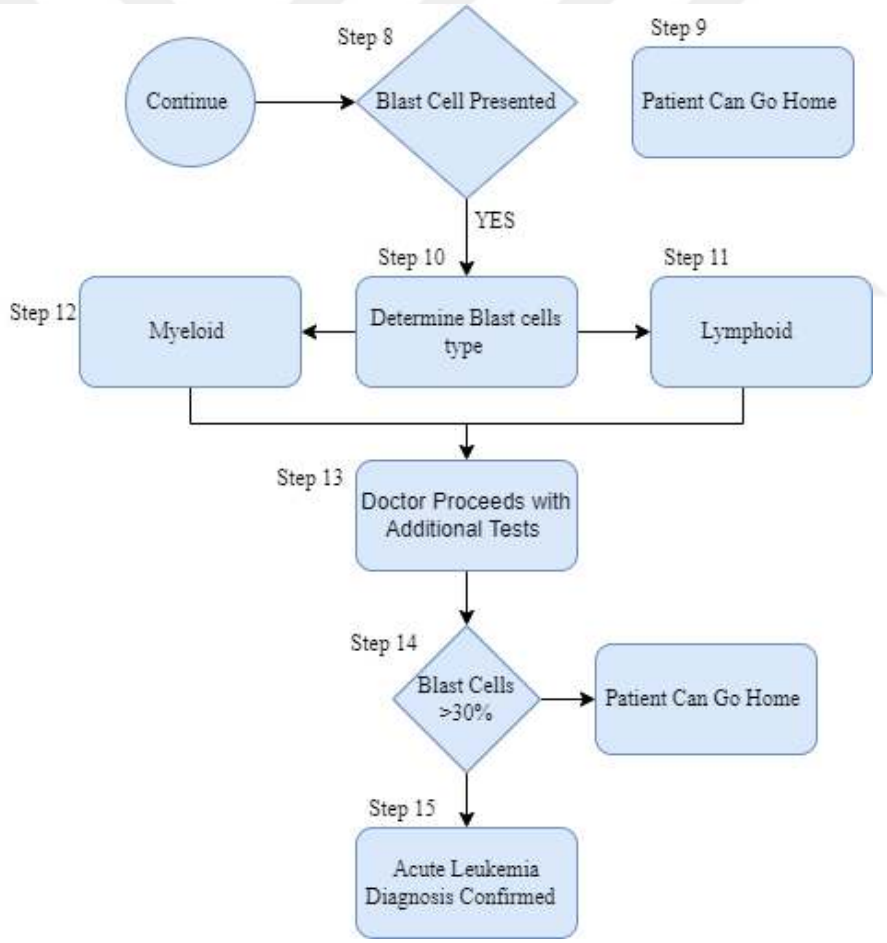

Figure 1.5. Steps to Confirm Acute Leukemia Diagnosis (Part 2)

1.5.1. Bone marrow exam Open pop-up dialog box

Additionally, doctors may detect persistent leukemia with a simple blood test before symptoms appear. If this occurs or has early signs of leukemia, the patient may be subjected to the following diagnostic tests.

1.5.2. A physical examination

Leukemia's physical indicators include pale skin anemia, lymph node swelling, and spleen & liver enlargement.

1.5.3. Blood testing

The physician can determine if an individual has leukemia by analyzing a blood sample to assess the presence of abnormal quantities of platelets or white or red blood cells. The blood test could potentially reveal the leukemia cells' presence; however, not all kinds of leukemia produce leukemia cells for circulation in blood. Leukemia cells may sometimes remain in the bone marrow.

1.5.4. Bone marrow examination

The doctor may advise you to have a bone marrow sample extracted from the hipbone. A large, thin needle is used to remove bone marrow. The sample is then submitted to a lab for examining leukemia cells. Explicit features of leukemia cells may be revealed by specialized testing, which will be utilized to assess treatment choices.

1.6. ML

ML is a renowned branch of Artificial Intelligence (AI), including mathematical relations and algorithms that were swiftly familiarized with clinical research. ML empowers computers to be programmed deprived of unambiguous experience and learn based on that experience. The aftermath of using these approaches in medical data processing has been astonishing and made noteworthy accomplishments in diagnosing disease (Ehrenstein et al., 2017; Wen et al., 2018). Research specifies that ML approaches significantly support difficult medical decision-making procedures in medical image processing by extracting and then examining the features of these images (J. Zhao et al., 2017). Since the number of diagnostic tools has grown and a large volume of high-quality data has been formed, the urge to have more advanced data analysis approaches.

Nevertheless, ML is appropriate for handling large amounts of complex data and might be an influential tool in perception and incapacitating disease (Al-Jarrah et al., 2015; Rajkomar et al., 2019). Recently ML algorithms have been presented to be comparable with experts in a diversity of tasks, from the initial diagnosis to prognosis approximation, along with an estimate of treatment difficulties. Though, many ML methods have continued to fail to establish their access to everyday clinical exercise because of various barriers and drawbacks. Figure 1.6 depicts ML algorithms.

Furthermore, to automate the AML and ALL detection process, it is advisable to utilize the power of modern ML algorithms. ML can be explained as an advanced version of the computer program that can automatically recognize specific problems in the data; in this case, it can precisely classify the ALL and AML. Consequently, ML uses an advanced computer algorithm to detect patterns within structured data. The ML is divided into two categories: unsupervised & supervised (X.-D. Zhang, 2020).

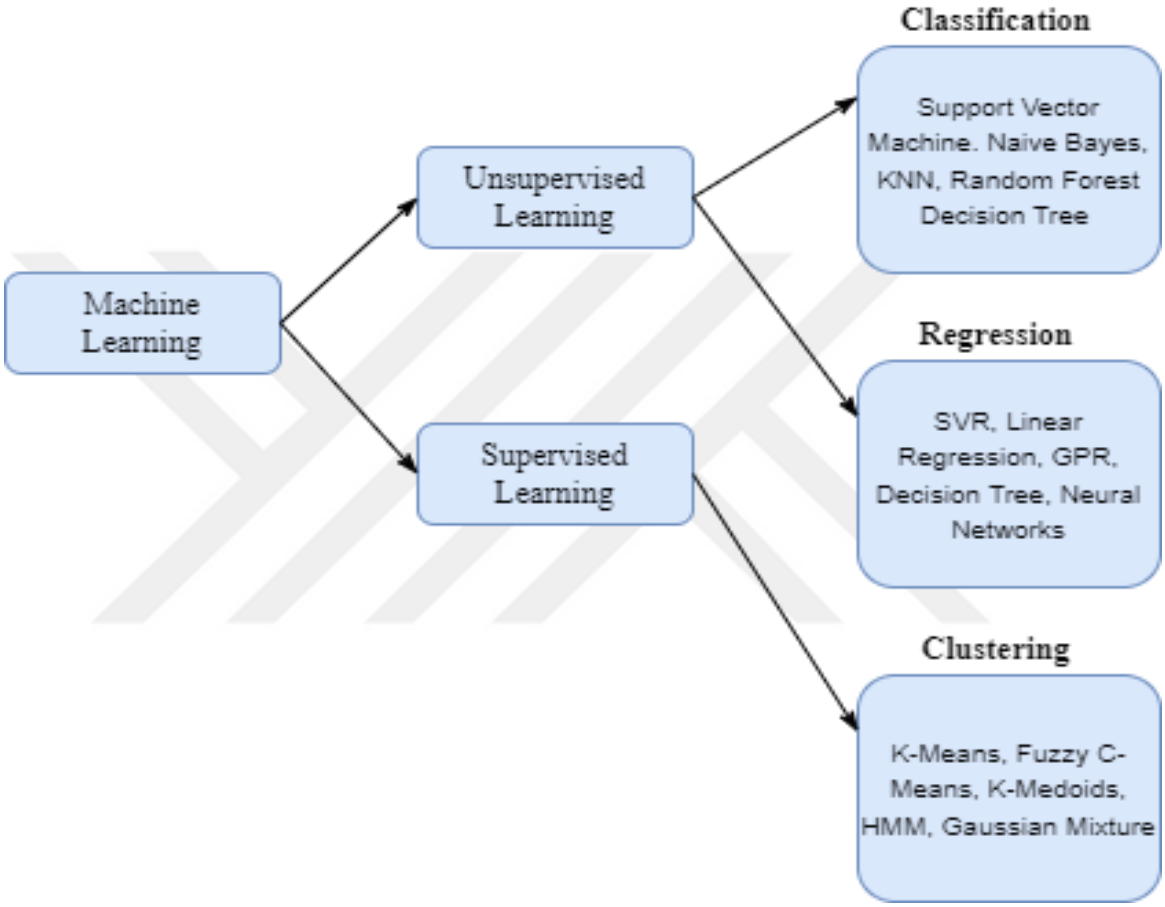

Figure 1.6. ML Algorithms

In unsupervised learning, the model identifies the different underlying structures in data without any instruction or user intervention. A typical unsupervised learning model is depicted in Figure 1.6. In contrast, supervised learning creates a prediction model with the help of a labeled dataset, completed with human help. A typical ML model is built in two phases. i.e., training and testing. An ML model generalizes a prediction equation from labeled samples in the training phase. The model's ability to correctly identify the unseen sample is tested during the testing phase. In simple Scenarios, such a predictive approach can be hand-programmed using a sequence of the if-then rules. Implementing these simple rules for every potential combination of variables in a complicated task such as microscopy is challenging; the code would become complex and untidy. Luckily, this complexity can be avoided by employing the ML algorithm, which can handle

the data with massive variables; meanwhile, writing rules for each possible variable's combination is unnecessary.

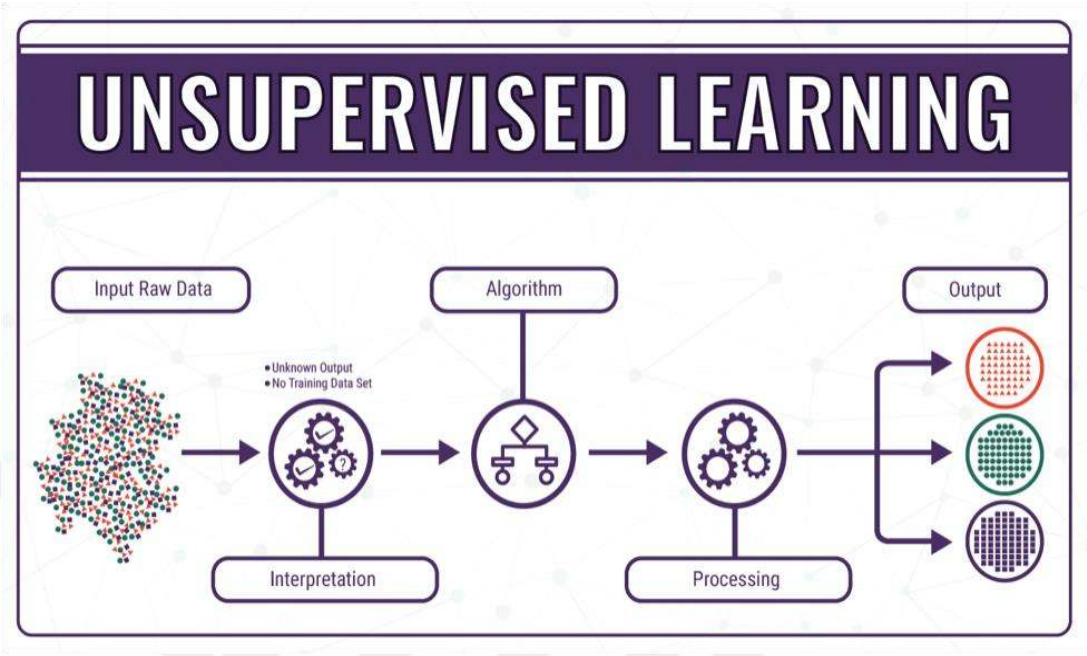

Figure 1.7. Typical cycles of Unsupervised Learning (Loon, 2018)

In supervised learning (Sammut & Webb, 2010), the proposed technique feeds each sample to the model as an input vector, producing a prediction for the output vector. After that, it matches the output to the label of the example. The label has often been called Ground Truth (GT). If a resultant prediction differs from GT, the algorithm adjusts the model's parameters to generate better forecasts in the following training step. A typical supervised learning model is depicted in Figure 1.7.

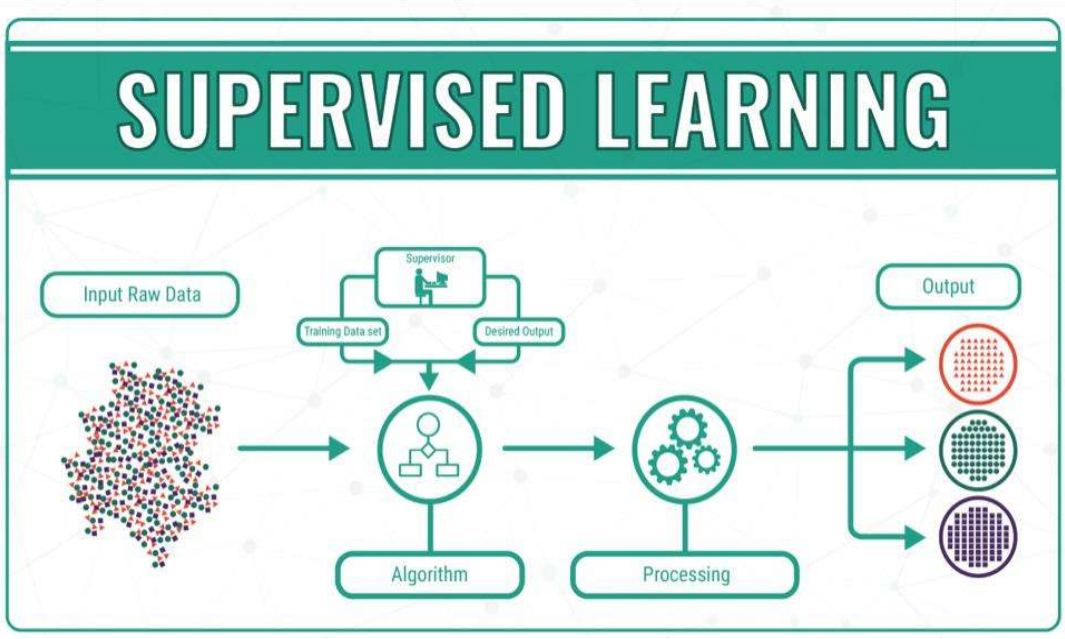

Figure 1.8. Typical cycles of supervised Learning (Loon, 2018)

1.7. Deep-RL

The category of AI & ML in which intelligent machines could be trained from their acts, in the same manner, people learn from their experiences, is known as Reinforcement Learning (RL). The individual has been rewarded or punished depending on their activities, which is a basic idea in this form of ML. Actions that lead to the desired results have been awarded (reinforced). A machine learns via trial and error, making this technique appropriate for dynamic conditions that are constantly changing. While RL is already present for years, it was later paired with DL, which produced spectacular results. A "deep" part of RL relates to several artificial neural networks and (deep) layers that replicate the human brain structure. DL needs vast training data and a substantial computer capacity. Over the past several years, data quantities have risen while processing power prices have decreased, allowing for an expansion of DL implementations. A well-known Go grandmaster lost to DeepMind's AlphaGo (Gibney, 2016), which used Deep-RL. This issue made many people think about how it could be used for other things. Go isn't the only application where Deep-RL is performed well. It is also being utilized in IoT-based real-time automation applications and performing excellently. Figure 1.9 depicts Deep-RL.

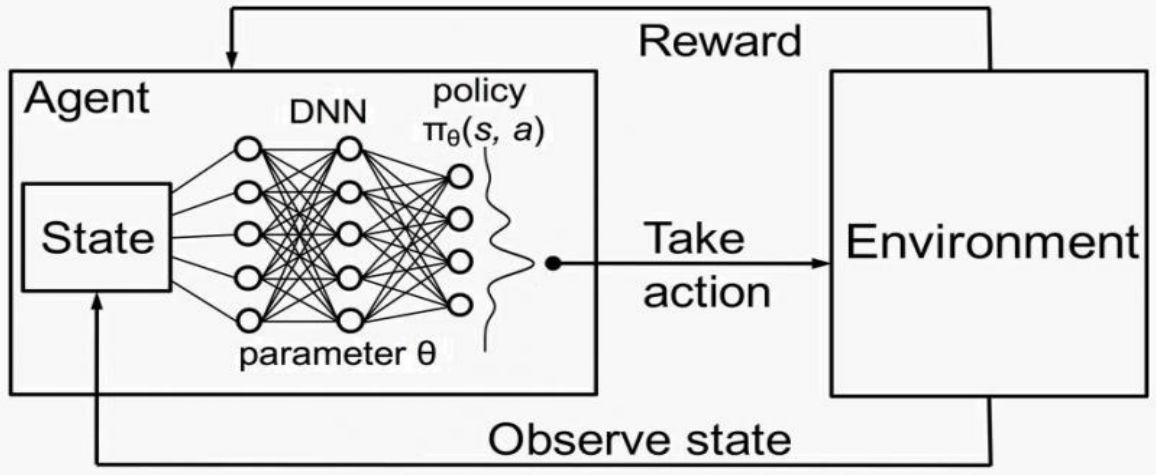

Figure 1.9. Deep-RL (Koo et al., 2019)

1.8. Problem Statement

Cancer is labeled as one of the deadliest global problems by the World Cancer Report because it destroys a significant portion of the world's population. According to (Mignogna et al., 2004), by 2025, a projected increase of 25 million cases can be seen. Moreover, the World Health Organization reported 14 million new cancer cases annually (Stewart & Wild, 2014). This ailment can substantially cause 8.8 million deaths annually (Boyle et al., 2008). Health practitioners, cancer has been testified as one of the lethal sicknesses of the genome. It has been the pivot of research by doctors, pathologists, biologists, data scientists, and other life science and health professionals. It is a need of the hour to have an advanced automated method that can quickly identify cancer and its subtype, i.e., ALL and AML. This research aims to identify cancer subclasses using DNA Gene Expression.

A plethora of research exists in the paradigm of cancer classification as many experiments have used multiple methods; however, exploiting the gene expression and interpretation data is one of the famous and advanced methodologies. In a cancer disease study, the essential aim is to find genes that induce the cells to metamorphose into cancer diseases (Y. Zhang, 2019). One well-known method that shows the extent of cancer in cells and molecules in technology is using Microarray-based datasets. A typical microarray comprises roughly 30,000 genes, which are exposed to examination for the detection and classification of cancer (Bolshakova et al., 2005). Such Microarrays can be described as assorted samples of DNA, RNA, tissues, or protein (Golub et al., 1999).

1.9. Contribution of The Thesis

Existing research can be divided into four major categories: work utilizing traditional data mining methods (Bolshakova et al., 2005; Özcan Şimşek et al., 2019), work using DL methods (Karim et al., 2019; Mostavi et al., 2020; Patel, 2021; Sharma & Rani, 2017; Sun et al., 2019). The

third is work utilizing ML methods (Bhuvaneswari & Therese, 2015; Y. Chen et al., 2015; Hooshmand, 2020; Vanitha et al., 2015; Verda et al., 2019) and classification using deep reinforcement (Ali et al., 2018; Balaprakash et al., 2019; Liu et al., 2019; Simin et al., 2020; Xu et al., 2019; Y. Zhang, 2019). The Cancer Genome Altus Dataset (Tomczak et al., 2015) is a particularly well-known dataset frequently used in these research studies. However, most of the literature utilized one type of model or some combination of different models. Most of the available literature for comparison is theoretical. To the best of our knowledge, the optimal compassion of ML models, DL models, and RL models are still unknown or unavailable in one place. This thesis aims to explore and implement the ML, DL, and Deep-RL-based models and provide a comprehensive comparative analysis so a researcher or ML practitioner can find all the works and implementation in one place.

1.10. Research Questions

In this thesis, we will aim to provide answers to the following research questions:

- What are the ML techniques that have been applied for the prediction of ALL and AML using gene expression data?
- What are the DL techniques that have been used for the prediction of ALL and AML using gene expression data?
- What are the RL techniques that have been applied for the prediction of ALL and AML using gene expression data?
- Among the various ML techniques that have been applied for the prediction of ALL and AML using gene expression data, which methods have been demonstrated to be fast to deploy and easy to implement?

1.11. Thesis Outline

Chapter 1: Introduction

The purpose of this thesis is to review and analyze the various machine-learning techniques that have been applied for the prediction of ALL and AML using gene expression data. This chapter presents an overview of the research problem, the research questions, and the contributions of this study.

Chapter 2: Literature Review

This chapter reviews the existing literature on ML techniques for predicting ALL and AML using gene expression data. The focus is on identifying the similarities, differences, limitations, and potential directions for future research in this area.

Chapter 3: Research Methodology

This chapter outlines the research methodology and experimental design employed in this study. It also discusses the databases used for this study, including their sources and characteristics.

Chapter 4: Results and Analysis

The study's results are presented and analyzed in detail in this chapter. The discussion includes comparing the performance of different ML techniques and evaluating their strengths and limitations.

Chapter 5: Conclusion and Future Directions

This chapter summarizes the main findings of the study and discusses potential directions for future research in the field of ALL and AML prediction using gene expression data.

2. LITERATURE REVIEW

2.1. Data Mining Techniques

Research on data mining has significantly increased in recent years, particularly in the field of big data analysis, industry, education, and technology.(Herland et al., 2014). Data mining examines large datasets that can be employed for several gene expression data aimed at recognizing diverse gene classifications. But the progress in data mining research is a breakthrough in bioinformatics. Furthermore, data mining methods have been applied mostly in experiments and research.

A general approach is presented by (Golub et al., 1999) for cancer classification depending on the monitoring of Gene Expression through exploiting the DNA microarrays and testing on humans for acute leukemias classification. They discussed a classification technique that revealed the difference between AML and ALL without previous information on such classes. They also expressed that artificially generated class predictors could classify new leukemia cases. They provided 2 Gene Expression datasets for the research community to build and classify the AML and ALL, respectively.

In a study by (Özcan Şimşek et al., 2019), authors suggested BM25 term and frequency weighting measures with relevance frequency and inverse document frequency metrics to weight mutations-based genes. They observed that the more changes a gene had in patients having a specific illness and the fewer variations it had in other people. They evaluated the recommended representations on the classification of cancer type. They used The Cancer Genome Atlas (TCGA) Dataset (Tomczak et al., 2015). They employed the *tf-IDF* schemes and BM25 variants to apply several machine-learning approaches. Their findings demonstrated that BM25-tf-rf representation improves f-score values along with classification accuracy. They also acknowledged that data representation based on BM25-tf-rf showed a maximum accuracy of 76.44 % and an f-score of 76.95 %.

A study based on Singular Value Decomposition (SVD) (Abdi, H. 2007) was conducted to choose informative genes and decrease the redundant data. Information gain is also utilized to ascertain valuable features and improve classification performance. The classification method is employed for the chosen features in the final phase. They also conducted some experiments on a subset of features from existing datasets. The experimental outcomes demonstrate that the presented feature selection and dimension reduction improve classification performance relating to the area under Receiver Operating Curve (ROC) and the accuracy.

The research work presented by (Tarek et al., 2017) provided a novel ensemble scheme for Cancer classification based on gene expression. The methodology developed to lead to a swift and appropriate scheme that surpasses the ensemble scheme proposed by (Okun, 2011). Their work also addressed and prevailed over three drawbacks, i.e., screening more cancer

categories, improving result accuracy, and alleviating the influence of over-fitting, in this research work. Furthermore, the authors utilized K-NN (K-NN) classifier as a base element of the ensemble.

In a study (Bolshakova et al., 2005) provided a data mining framework that enabled various cluster validity and clustering techniques on DNA microarray data. They expressed that the Machaon CVE tool (K-Means Clusters) may help enhance the data analysis findings' performance and forecast the number of meaningful clusters within microarray datasets using Gene Expression Dataset (Golub et al., 1999). They presented that this systematic assessment may significantly improve genomic expression analysis for knowledge extraction applications. They concluded that their created software system might be utilized to verify and cluster applications other than DNA microarray/expression analysis. Table 2.1 shows the comparison of the different works.

Table 2.1. Comparison of different approaches based on data mining techniques

Reference	Method	Dataset
(Golub et al., 1999)	K-means clustering	Gene Expression dataset (Golub et al., 1999)
(Özcan ŞiMşek et al., 2019)	BM25-tf-rf	TCGA Dataset (Tomczak et al., 2015)
(Tarek et al., 2017)	Ensembler (KNN classifier)	Colon dataset (Alon et al., 1999)
(Bolshakova et al., 2005)	K-means clusters	Gene Expression Dataset (Golub et al., 1999)

2.2. Classification Using DL

The authors presented numerous Convolution Neural Network (CNN) approaches that use irregular inputs of gene expression to identify samples as normal or malignant (Mostavi et al., 2020). They built three CNN architectures depending on various designs of convolution, and gene embedding approaches 1D-CNN, the 2D-Hybrid-CNN, and the 2D-Vanilla-CNN. They evaluated the algorithms on 10,340 samples from (TCGA) (Tomczak et al., 2015), including 33 types of cancer & 731 matching normal tissues. Their models achieved good prediction accuracies of 93.9 to 95.0 % across 34 categories.

In (Karim et al., 2019) presented OncoNetExplainer, a novel technique for making understandable predictions of the cancer kinds using gene expression data. They utilized genetic data from the TCGA dataset (Tomczak et al., 2015) on 9,074 cancer patients representing 33 distinct cancer types for training VGG16 and CNN networks employing guided-gradient class activation maps. They also created heat maps to highlight relevant biomarkers and calculated feature relevance based on mean absolute effect to prioritize the top genes among all types of cancer. Their experimental results showed that both models

achieved a high degree of confidence in properly identifying cancer kinds, with an average accuracy of 96.25 %.

Another study (Patel, 2021) introduced a tumor detection system that used CNN on breast mammography images as an innovative way to improve training time and feature extraction. Their suggested model was learned and assessed by utilizing Digital Database for Mammography Screening (Bowyer, K., Kopans, D., Kegelmeyer, W.P., Moore, R., Sallam, M., Chang, K. and Woods, 1996) for saving training time. They randomly chose 1000 malignant and benign cases for training, with an additional 450 malignant and benign cases employed for validation. According to their experimental results, the CNN classifier was trained with the most recent training scheme to yield an 11% error in 150 epochs, whereas the validation range error was less than the 22%.

However, (Sun et al., 2019) presented deep genome learning, a Deep Neural Network (DNN) based technique for studying the link between genetic variants and characteristics. They examined 6,083 samples' Whole Exon Sequencing mutation records across 12 cancer types available from TCGA (Tomczak et al., 2015) and WES files from healthy examples of 1,991 obtained from the one thousand Genomes project. They built twelve specific models to discriminate between healthy and cancer cells, a clear strategy that could identify both cancer and healthy tissues, and a mixed model, which could distinguish among all twelve cancer types based on Genome Deep Learning (GDL). They showed that accuracy of mixture & total specific models for cancer detection was 97.47 %, 70.08 %, and 94.70 %, respectively.

In their study (Sharma & Rani, 2017) presented an efficient framework for cancer categorization using genetic algorithms and DL. They built a novel framework for scaled computing on a previously developed ML-based method named H2o. Their suggested technique had the critical benefit of scalable cancer categorizing utilizing a Gene Expression Dataset (Golub et al., 1999) and automatic extraction of features employing DL. They used their approach to categorize cancer kinds based on publicly accessible GE data. They also used parameter tuning utilizing the genetic algorithm to increase classifier performance. They claimed that their proposed method achieved an accuracy of 92%. Table 2.2 shows the comparison of the different works along with their achieved accuracies.

Table 2.2. Comparison of different approaches based on DL methods

Reference	Method	Dataset	Results (%)
(Mostavi et al., 2020)	1D-CNN, 2D-Hybrid-CNN, and 2D-Vanilla-CNN	TCGA Dataset (Tomczak et al., 2015)	93.9-95.0
(Karim et al., 2019)	VGG 16	TCGA Dataset (Tomczak et al., 2015)	96.25
(Sun et al., 2019)	Genome Deep Learning (GDL)	TCGA Dataset (Tomczak et al., 2015)	97.47
(Sharma & Rani, 2017)	CNN	GeneExpression Dataset (Golub et al., 1999)	92.00

2.3. Classification Using ML Methods

Currently, the prediction of leukemia in its early stages is quite demanding for physicians, and it can be performed by applying computer-aided disease diagnosis tools in medical areas (Mallick et al., 2020). A diverse range of ML methods was utilized for medical datasets for preparing an intelligent diagnosis structure. As a result of the digital revolution and progress in information technology, tremendous data is created out of medical sectors. The ML methods are appropriate for analyzing such large data, and also numerous methods were assumed for diagnosing diverse diseases (Mahmood et al., 2020). Different diagnosis data are obtainable through medical records at multiple hospitals for effectively running any ML algorithms; microarray technology generates a large quantity of DNA expression data from such hospitals. Classification of these data are more indispensable for the diagnosis at an early stage and is also responsible for any genetic disease. Research on gene expression data is amongst the most famous regions in ML –based data analysis. Diverse ML algorithms were used for data analysis.

A microarray gene expression-based technique is presented by (De Guia et al., 2019). They used ML techniques named Gradient Boosting (GB) alongside SVM to identify cancer genes. They provided the study of genetic factor expression data & categorized the findings based on the dataset of Gene Expression. They claimed that GB had a greater precision of 64.71 %. Another study by (Alamgir Sarder et al., 2020) proposed five feature selection strategies to discover the most important genes and six different algorithms to categorize & analyze cancer diseases. They obtained data about leukemia cancer from the Kent Ridge Biomedical Data Repository in the United States. As a t-test, they utilized five FSTs: RF, Boruta, Wilcoxon sign rank-sum, least absolute shrinkage, and selection operator. They employed a simulated dataset to test the validity of the suggested strategy. Their experimental results showed that integration of FST based on LASSO and NB classifier yields a maximum of 99.95 % classification accuracy.

But, the study (Hooshmand, 2020) created an innovative ML model against cancer approaches that provide excellent outcomes with perfect specificity and sensitivity.

They employed complete genome sequence data from the next-generation RNA sequencing methods in Genotype-Tissue Expression and TCGA studies for healthy and cancerous tissues. Their suggested approach to data manipulation, along with generic algorithms,showed flawless classification. They observed that precision, specificity, and sensitivity were equal to one for most tumors, even with little data. They claimed that their suggested method showed 90 % precision.

Another research work by (Verda et al., 2019) assessed the efficacy of a Logic Learning Machine (LLM) in categorizing cancer patients utilizing a collection of eight publicly accessible gene expression datasets collected from the GEO repository cancer diagnosis. They determined LLM accuracy using a summary ROC curve assessment and area under a ROC curve. They compared the proposed method with the typical supervised approaches such as Decision Tree (DT), Artificial Neural Networks (ANN), SVM, and K-NN classifiers in cross-validation. Their experimental results showed that LLM beat all other methods except SVM with an accuracy of 99%.

Authors used t-Statistics to identify specific genes (S. Khan et al., 2019). They differently expressed 746 genes & 186 genes for the 2000 colon dataset and 7128 of the leukemia dataset (Gene Expression Dataset (Golub et al., 1999). They determined prediction accuracy using 10-fold cross-validation & three distinct machine-learning algorithms. They deployed the ML techniques 100 times to get average forecasting accuracy. In their work, the authors also utilized a box plot for comparison. Their finding indicated that the three approaches' patient estimation accuracy had more than 90% and 87% throughout the examination of the Colon and Leukemia datasets, respectively, and RF showed superior prediction accuracy.

Yet another work by (Chen et al. 2015) used SVM to investigate the 1,760,846 patterns of somatic mutations detected from 230 255 patients with cancer and gene function info. They employed the COSMIC dataset in their work. They ran a classifier experiment across 17 tumor locations, utilizing the available gene route, gene symbol, somatic mutation, and chromosome as indicators for 6,751 patients. They observed that baseline accuracy was 57% using gene characteristics but improved to 62% after including chromosomal and mutation information. They claimed that the forecast of 5 primary spots, such as the large intestine, pancreas, liver, skin, and lung, had an F-measure higher than 70% amongst the predicted main tumor spots. They showed that the big intestine model got first with an F-measure of 87%.

Nonetheless (Devi Arockia Vanitha et al., 2014) described an efficient technique for gene categorization based on SVM. They expressed that SVM was the supervised learning technique that could solve complex classification tasks. They employed mutual information between the class and genes label f for finding informative genes. They used chosen genes to train the SVM classifier, & its testing capacity was evaluated using the

Cross-Validation technique. Their suggested technique performance was assessed using a microarray Gene Expression Dataset (Golub et al., 1999). They stated that SVM beat other ML algorithms with 67.74 % accuracy. Table 2.3 shows the comparison of the different works along with their achieved accuracies.

Table 2.3. Comparison of different approaches based on ML methods

Reference	Method	Dataset	Results (%)
(Mallick et al., 2020)	DNN	Gene Expression Dataset (Golub et al., 1999)	98.20
(Mahmood et al., 2020)	RM, GM, C5.0 DT	(Pan et al., 2017)	87.40
(De Guia et al., 2019)	GB and SVM	Gene Expression Dataset (Golub et al., 1999)	99.90
(Devi Arockia Vanitha et al., 2014)	SVM	Gene Expression Dataset (Golub et al., 1999)	67.74
(Verda et al., 2019)	ANN, SVM, and KNN	GEO repository	99.00
(S. Khan et al., 2019)	SVM, LDA, and RF	Gene Expression (Golub et al., 1999)	90.00

2.4. Classification Using Deep-RL

RL a peculiar class of ML, has also built applications in cancer and medical oncology research regarding uncovering the optimal treatment strategies and a computer-aided diagnosis of disease. In a typical RL model, an agent, for instance, uses the physician's interaction with the environment to accomplish an objective based on the result that the physician desires to improve (reward). The learning procedure of an agent in an RL is a continuous process. However, the environmental interactions occur at a discrete time. Once an environment's condition is acquired, the agent picks a specific action to interact with. Then the environment responds to the action taken by the agent, and in return, the reward that the agent like or is unlikely to obtain is finally settled (Sutton & Barto, 2015).

A combination of ANN and RL techniques is utilized to diagnose and classify melanoma, squamous & basal-cell skin cancers (Simin et al., 2020). They pre-processed the cancer photos from the ISIC site/archive and retrieved ten characteristics. After this, they sent retrieved characteristics into Neural Network (NN) for classification. They used a well-known deep-RL approach-solving methodology called the Q-learning technique to achieve this goal. Their outcomes were compared with the test networks and revealed the preferable performance of the RL method for selecting the optimum neurons within hidden layers. They showed that the algorithm achieved 93.78 % accuracy using the RL technique.

The research work (Xu et al., 2019) described a unique Deep-RL strategy for breast cancer categorization. They developed Recurrent Neural Network (RNN) to make classification judgments and forecast the picture area's position to be discovered at the next sampling interval. They used reinforcement techniques to improve the entire network to acquire

an optimum strategy for classifying the region because the region selection procedure was non-differentiable. They only needed to analyze a portion of pixels inside the raw picture using their new look, organize, and investigate technique, resulting in considerable savings in computing resources without losing speed. Their approach achieved 96% accuracy using the public breast/cancer histopathology dataset.

Furthermore, (Balaprakash et al., 2019) created an architecture based on Deep-RL to automate the prediction models based on DL for a relevant cancer data class using the NCI-ALMANAC database. We provide bespoke building blocks that enable domain specialists to comprise specific features of cancer data. They also validate that their technique discovered DNN topologies with much lesser parameters, faster training times, and accuracy comparable to or greater than manual architectures. They investigated their technique's scalability on approximately 1,024 Intel-Knights-Landing units of Argonne Leadership Computing Facility's Theta supercomputer. They claimed that their suggested method showed an accuracy of 98.5%.

Table 2.4. Comparison of different approaches based on Deep-RL

Reference	**Method**	**Dataset**	**Results %**
(Simin et al., 2020)	Deep-RL / Q-Learning	ISIC site/archive	93.78
(Xu et al., 2019)	Deep-RL / Q-Learning	Public breast/cancer histopathology database	96.00
(Balaprakash et al., 2019)	Deep-RL / Q-Learning	NCI-ALMANAC database	98.50
(Ali et al., 2018)	Deep-RL / Q-Learning	LUNA's LIDC/IDRI database [29]	64.40
(Y. Zhang, 2019)	Deep-RL	FCTM (Di Cataldo & Ficarra, 2017), EECO (Im et al., 2016), and Thydeep (Jiang et al., 2017)	86.69

Furthermore, authors (Ali et al., 2018) presented an ANN-based Deep-RL approach for the timely identification of lung cancer using CT images. AlphaGo motivates the work. Their DL method took the raw CT images as input, perceived it as the state's collection, and outputs nodule presence classification. They trained their model by using LUNA's LIDC/IDRI database (Armato et al., 2011). Their training results provided a 99.1% overall accuracy. The experimental results of their test produced a 64.4% accuracy. Another research work by (Liu et al., 2019) showed some exemplary Deep-RL techniques that might be used to identify lung cancer. They described the most prevalent kinds of lung cancer and the essential features of every type. They highlighted open obstacles and potential future research paths of using deep-RL to identify lung cancer, which was predicted to support the development of smart healthcare with the medical IoT.

Although (Y. Zhang, 2019) offered a unique Deep-RL approach for thyroid cancer identification and prediction to imitate the visual attention process. They discussed that the Sal-deep network added the saliency map to the Deep-residual network, which increased the feature recovered from various areas based on the mask map. They expressed the Sal-deep network was a network paradigm that may operate successfully for the benchmark with multiple data sets. They claimed that sal-deep networks enhanced network complexity while improving network efficiency. They showed that their proposed technique showed a precision of 86.69 %, susceptibility of 89.25 %, and specificity of 86.91 % on FCTM (Di Cataldo & Ficarra, 2017), EECO (Im et al., 2016), and Thydeep (Jiang et al., 2017) datasets. Table 2.4 shows the comparison of the different works along with their achieved accuracies.

3. MATERIALS AND METHODS

3.1. Materials

The fast development of cancer genome data sets coincided with the rapid expansion of analytical techniques for genome studies analysis of microarray. In classifying leukemia cancer datasets, several researchers have tried a diverse range of classification functions that have been presented. The data in our work is initiated from proof-of-concept presented by (Golub et al., 1999). This proves that gene expression monitoring might be used to categorize cancer and suggests a way to determine novel cancer classes and allocate them to previously known cancer classes. We have utilized microarray data (Golub et al., 1999), and raw data is obtained from scanning genomic microarrays out of 72 individuals. These individuals have one of the cancer categories under consideration, such as (leukemia). This thesis corresponded to 7129 gene values for every patient that has contributed. In this study, ALL contributes 65%, AML contributes 35% of the overall dataset, and gene description and accession number are contained within this section. In this study, we used a gene expression dataset and applied various classification techniques, including supervised ML methods such as LR (Joachims, 1999), SVM (Mukherjee, 2003), K-NN (Fix & Hodges, 1989), DT (Svetnik et al., 2003), RF (Mallick et al., 2016), XGBoost (Kashef et al., 2020), Gaussian NB (Begum et al., 2016), and NN (Kocatürk et al., 2022). We also employed a Deep-RL method. The proposed methodology for this study is illustrated in Figure 3.1.

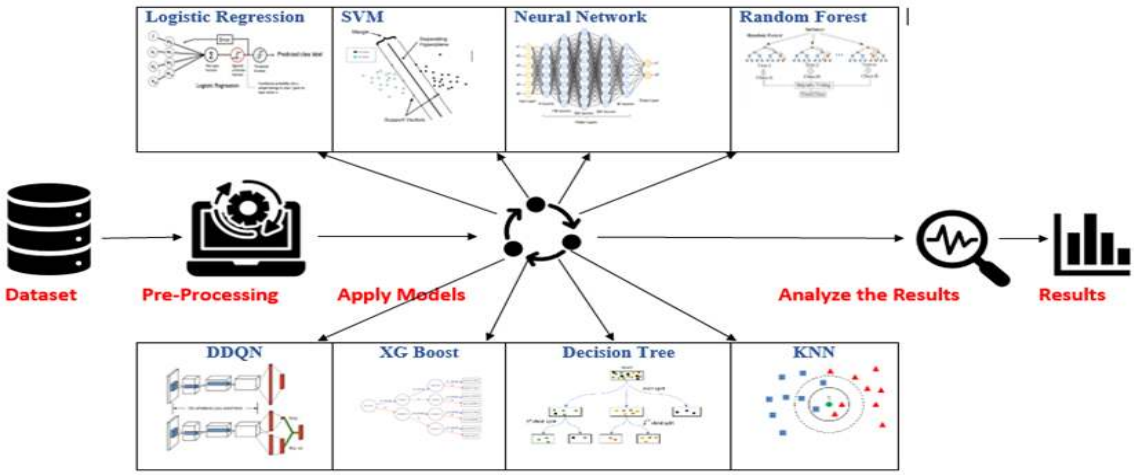

Figure 3.1. Proposed approach diagram

3.2. Microarray Technology

Microarray technology has been extensively utilized in gene quantification technology for two decades. Microarray technology is still considered better than RNA-seq technology because of its low cost. Besides, the microarray is regarded as the first technology to measure

expression levels of all genes simultaneously. The most utilized platforms for microarray are Illumina and Affymetrix (Göhlmann & Talloen, 2017; Illumina, 2009). The expression values are obtained through microscopic DNA spots connected to a solid surface and follow a hybridization procedure. After the completion of this process, then, with a laser, the gene expression values are measured (Taqman, 2009; Zahurak et al., 2007). Lastly, the quantification levels are measured and then exported to a CEL file (Schena et al., 1995).

Furthermore, The DNA, RNA, and protein samples on the microarrays represent the human genome (Cook et al., 2016). The sample will determine the type of microarray on the slide, such as RNA, DNA, or protein combinations. The available cancer genomic data makes it feasible to study cancer gene expression. Most data sets were provided for lung or breast cancer, while others had sample sizes of less than a hundred people. A breakthrough in gene expression profiling, microarray profiling, has been mostly employed to analyze gene expression in cancer (Tarca et al., 2006). Methods of cancer classification, as well as their assessment and accuracy.

3.3. Dataset

The dataset demonstrated how new cancer cases might be identified using gene expression and offered a generic technique for finding novel cancer classes and classifying tumors. These data were analyzed to determine the classification of individuals with ALL and AML. The dataset that the system will run is defined. Then a series of actions to be performed on the dataset is defined amongst loading and connecting to the dataset repository. In this work, we used leukemia gene expression data which was collected by (Golub et al., 1999).

Additionally, the proposed framework has been tuned and tested against the Leukemia cancer dataset to lift the assurance in utilizing the proposed system. The dataset contains 72 bone marrow samples, including 38 B-cells, 09 T-cell ALL, and 25 AML samples. In the original study, the data from 37 patients were utilized for training, and the rest from 35 patients was used for testing. But this study divided the testing dataset into two groups with initial 38 training samples and 34 test samples, along with measurements corresponding to ALL and AML samples fetched from bone marrow and peripheral blood, respectively. The dataset shows how new cancer cases could be classified by gene expression monitoring (via DNA microarray). It provides a general approach for identifying new cancer classes and assigning tumors to known classes. These data were used to classify patients with AML and ALL. Besides, intensity values have been rescaled so that broad powers for every chip can be equal.

Moreover, corresponding to the dataset determined in the gene expression dataset, the preprocessing module operates the dataset. The formulation comprises thresholding, logarithmic transformation, filtering, and data normalization. Those measures are crucial to be performed formerly to the actual classification occurring.

3.4. Data Preprocessing

3.4.1. PCA-based feature selection

In the dataset, removed few columns features were form utilized datasets through preprocessing technique because those deleted columns do not have any effect on the final predicted outcomes of proposed models. Thus, features have to be reduced with important features adept at discriminating classes to reduce the complication of classification, decrease computational requirements, and enhance the prediction of classifiers. This sequentially improves the accuracy of type, simplifies data visualization, and offers a better grasp of the fundamental data by disclosing the interrelationship between features.

Since there are 38 patients' rows in the training set and the other 34 rows in the testing set, each dataset has 7129 gene expression features. But we haven't yet associated the target labels with the right patients. We will recall that all the labels are stored in a single data frame. Let's split the data so that the patients and labels match up across the training and testing data frames. Undoubtedly, there is some variation in the scales across the different features. Many ML models work much better with data on the same scale, so let's create a scaled version of the dataset.

Moreover, there is a degree of variety in scales across numerous characteristics. This work created a scaled version of the dataset as multiple ML models perform superiorly with data using the same scale. Figure 3.2 presents the visualization of data before and after scaling. After scaling, the data is clear, and the values line is easy to visualize.

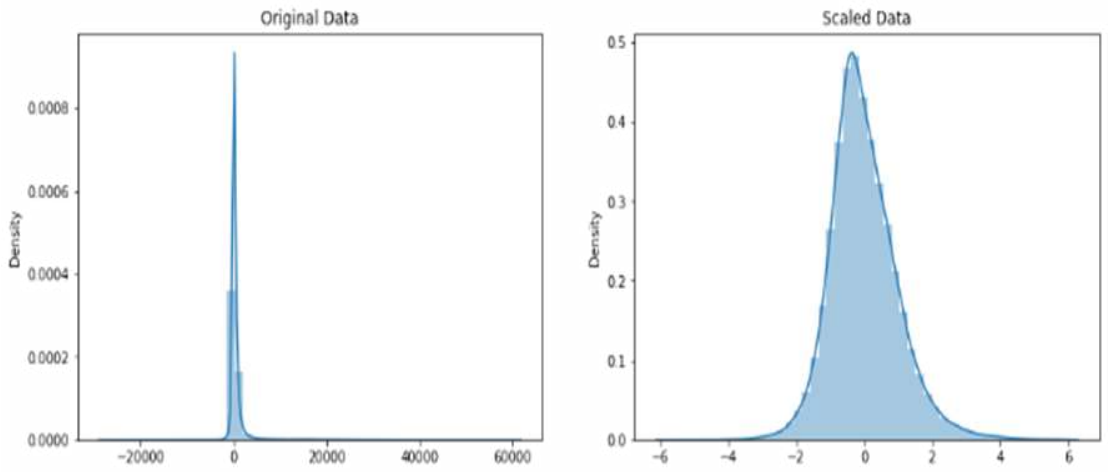

Figure 3.2. The visualization of data before and after scaling (PCA)

3.4.2. RF Importance

This work also utilized the RF importance feature selection. Since there are 72 patients, combined training, and testing sets, each of whom is labeled either "ALL" or "AML," depending on the type of leukemia they have. So, the value of ALL is 47, and for AML, the

value is set to 25. Also, each of those datasets has 7129 gene expression features. The setting is like the PCA version, with 38 patients' rows in the training set and the other 34 rows in the testing set.

Moreover, the goal is to create a scaled version of the dataset as numerous ML models perform superior with data using the same scale. Figure 3.3 illustrates the visualization of data before and after scaling.

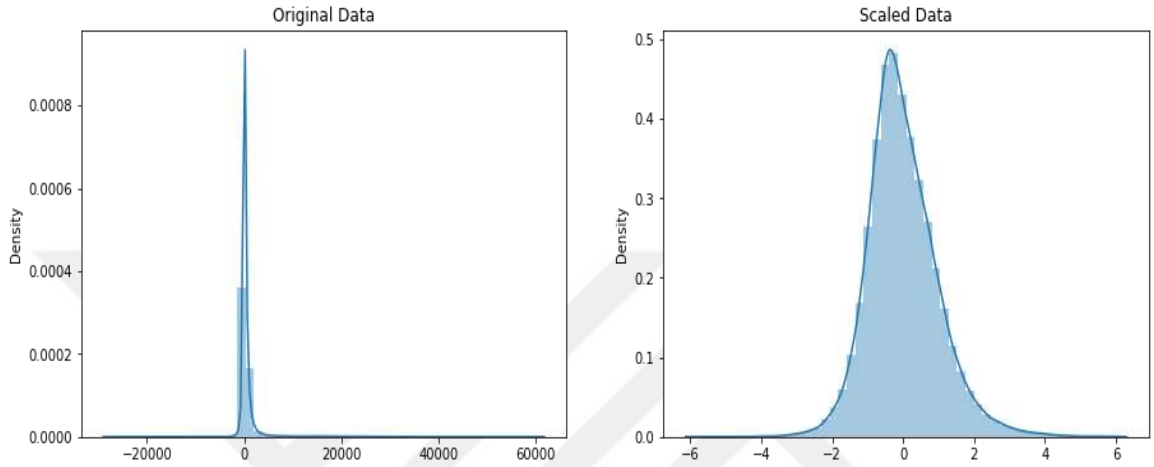

Figure 3.3. Data before and after scaling (RF importance)

3.4.3. LASSO Regularization-based feature selection

The last version of the feature selection algorithm is Lasso Regularization to select the important features. Similar to both previous feature selection algorithms. There are 72 patients, combined training and testing sets, each labeled either ALL or AML, depending on their leukemia type. The value of ALL is 47, and for AML, the value is set to 25; each dataset has 7129 gene expression features. The setting is similar to PCA and RF importance versions, with 38 patients' rows in the training set and the other 34 rows in the testing set. Figure 3.4 shows the visualization of data before and after scaling.

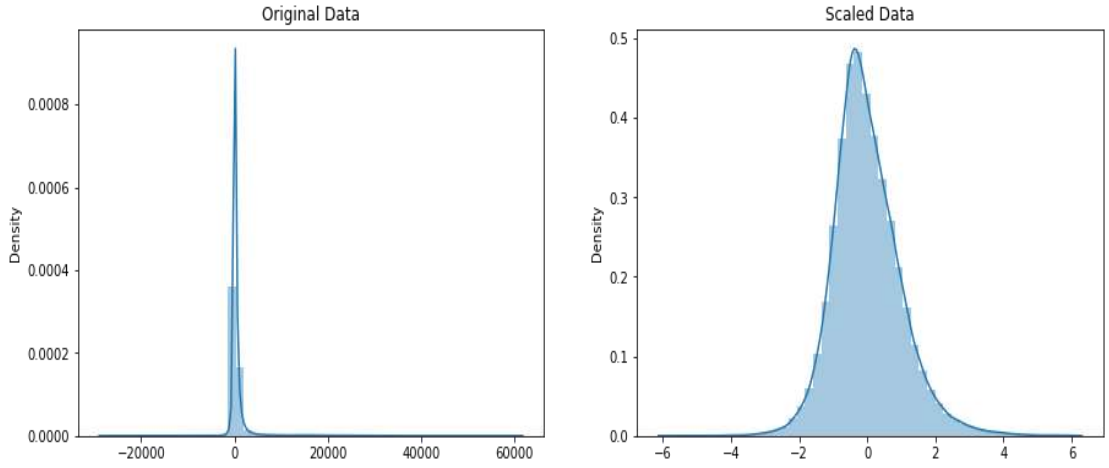

Figure 3.4. Data before and after scaling (Lasso Regularization)

3.5. Feature Extraction

The significant object in feature extraction is to choose appropriate features, with sufficient data regarding input to distinguish the different classes effortlessly. The most advantageous solution is choosing relevant features by hand. Though, for this instance, it would not be feasible to handle that. To reduce the dimensionality of the dataset, this study used Principal Component Analysis (PCA) for the feature selection (Kocatürk et al., 2022). This is referred to as a data reduction strategy in most cases. The PCA can allow you to choose the number of dimensions or primary components that appear in the converted outcome. It will be used to determine the relative relevance of the features. We will break down the information into two categories.

Furthermore, PCA is used for reducing the dimension of the data. The maximum no. of PCA components is related to the rank of the covariance/correlation matrix. Having a data frame with shape *[n_rows, n_cols],* the covariance/correlation matrix would be *[n_cols, n_cols]* with a maximum rank equal to the minimum *[n_rows, n_cols].* Thus, we can have substantial PCA components at most minimum *[n_rows, n_cols]* due to the maximum covariance/correlation matrix rank. Besides, in this work, the dataset has dimensions (38, 7129). So, the maximum no. of PCA components is 38. Figure 3.5 shows the dataset that applied to the PCA and the PCA Graph.

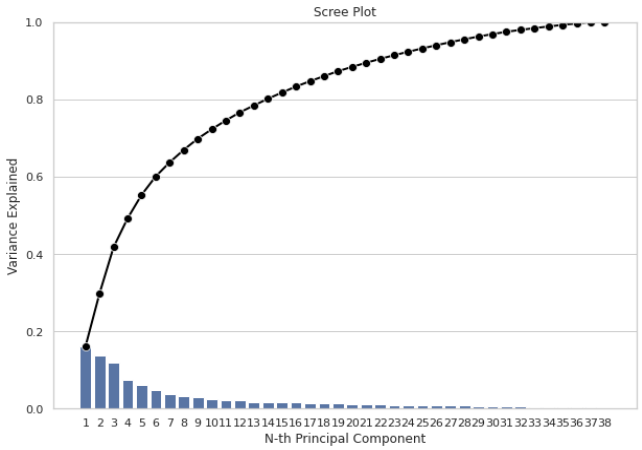

Figure 3.5. PCA Graph

Besides, we applied six options (For all our Algorithms), such as the minimum number of Components (2), the maximum number of Components (38), and four other values between minimum and maximum.

Moreover, in this work, we have utilized two other feature selection algorithms, Lasso Regularization (Hamada et al., 2021) and RF Importance (Huljanah et al., 2019). The settings of these algorithms are the same as the PCA version.

3.6. Classification Algorithms

Classification is one key method of data mining. One of the valuable features of these methods is to find concealed information. Dataset classification is in two-steps practicability, where a classification model is built up called a learning step and a classification step where the model is used to foretell class naming for giving dataset (Han & Kamber, 2012).

3.7. Classification using ML Algorithms

3.7.1. Classification using LR

LR can be described as a statistical technique that uses a logistic function to epitomize a variable (Wright, 1995). Similar to the original logistic model, a binary logistic model has a dependent variable with binary values, denoted with the labels "*0*" and "*1*". For instance, determining if a sample contains a specific class. Each detected type is assigned a probability of "*0*" to "*1*", while the sum of the probability of both classes will be *1*.

The logistic model measures the probability of a given class or event, such as healthy/sick. The logistic model's log chances for the value labeled "1" are a linear combination of independent variables that can be binary or continuous. The log probabilities are called logit, derived from the logistic unit in the literature. The equation of LR can be explained as follows:

$$\log\left(\frac{Q}{1-Q}\right) = \alpha_0 + a_1 X \tag{3.5}$$

Where Q denotes the probability and $\frac{Q}{1-Q}$ denotes the likelihood of the results. The figure below shows the graphical representation of LR. However, the working technique of LR is similar to the Linear Regression model. Still, the LR model uses a more complicated cost function, which may be defined as the 'Sigmoid function' or the 'logistic function' rather than a linear function (L. Zhao et al., 2001). Figure 3.6 illustrates curves of Linear Regression compared to LR.

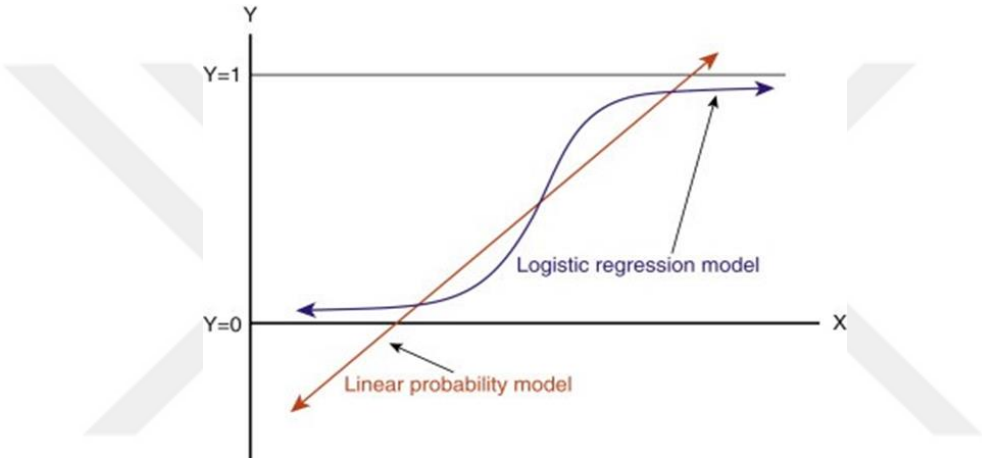

Figure 3.6. Comparison of linear and logistic curves (Mehmood, Asad, 2019)

3.7.2. Classification using a SVM

The SVM is a supervised learning method commonly employed in ML -based applications (Joachims, 1999). The SVMs have supervised learning models related to learning algorithms that examine data and identify patterns and are also utilized for regression analysis and classification tasks. According to a group of training samples, every marked as an association to one of two classes, an SVM training algorithm generates a model that allocates new samples within one class or the other, which causes it to be a non-probabilistic binary linear classifier. Moreover, new samples are mapped into that space and forecasted to pertain to a class based on which part of the gap samples fall on. In short, SVM is a binary classification process that distinguishes a group of data points from another. SVM is amongst the most often applied techniques in classification. The SVM and its variation are currently one of the top eminent classification methods with computational perks contrary to competition (Cristianini & Shawe-Taylor, 2000). The SVM has demonstrated functional benefits in tumor classification because of SVM provides fast computation, easy implementation, statistical proficiency along with capability to perform gene selection. Furthermore, it utilized numerous categories of data

in a large-scale dimension completely with a small number of examples and a substantial number of genes.

Consider a situation as follows: dataset D contains a series of samples $a1, a2, ap \in A$ with corresponding labels $y1, y2, yn \in y$. A separating hyperplane classifier is to train a function $f: M \rightarrow n$ from D that is exploited to forecast the label for each sample $a \in A$ by $f(a)$ and characterized into corresponding categories as $y \in \{-1, +1\}$. $A1, +1$ using the untying hyperplane classifier technique. In this case, the condition $W^T A + b > 0 \, for \, yi = 1$ known as the separation hyperplane property (Chih & Chih, 2022). Figure 3.7 portrays an illustration of SVM.

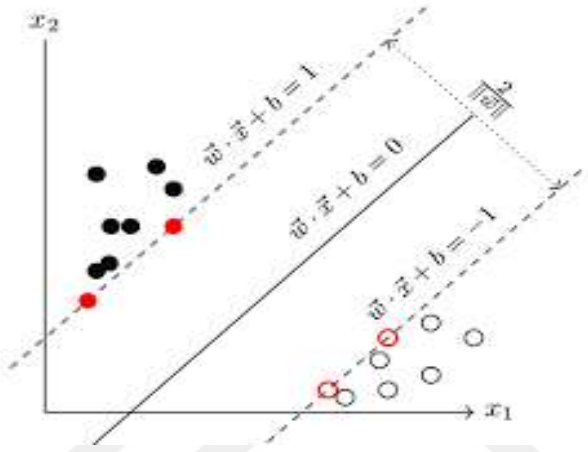

Figure 3.7 SVM Classification curve (Awad & Khanna, 2015)

3.7.3. K-Nearest Neighbors (K-NN)

The K-NN can be described as a non-parametric classification technique developed by Fix and Joseph Hodgesin in 1978 for applying classification issues (Saini et al., 2013). This method imposts the data regarding its neighbor points for output label classification (Newman & Sproull, 1974). This algorithm is extensively utilized for pattern recognition along with classification applications. This algorithm can predict the target class more accurately and unpretentiously. The K-NN is a lazy-learning algorithm with its function predicted only locally, and the entire calculation is belated until the classification process (Mejdoub & Ben Amar, 2013).

The K-NN algorithm utilized in our work considers the output as a target class. The value of K is engaged as a real-valued positive integer (Cunningham, 2007). The KNN algorithm is subtle toward the local part of the input, which makes the algorithm further exclusive to the classification issue. Further steps regarding the algorithm are as follows:

- Estimate Euclidean distance $d(x, x_i)$ where $i = 1, 2 \ldots n$ amongst the points.

- Sort above attained n distances in ascending way.
- Take k as a real-valued positive integer, and then the first k distances are obtained from the above structure.
- Use these k distances and calculate k points.
- For $k \geq 0, k_i$ is the number of points resembling the i^{th} the group within k points.
- Then check the state of k_i, k_j on the condition that the value of i is not equal to j, keep x in the class i.

Here the value of k is taken about the input; k turn into a larger number, then it will decrease irrelevant features' impact on the output category. The presence of irrelevant features in the input dataset will decrease the K-NN algorithm's precision despite its elevated classification efficiency (Frigui & Gader, 2009). Therefore, feature selection and pre-processing are important for the proposed work. Figure 3.8 illustrates the K-NN classification.

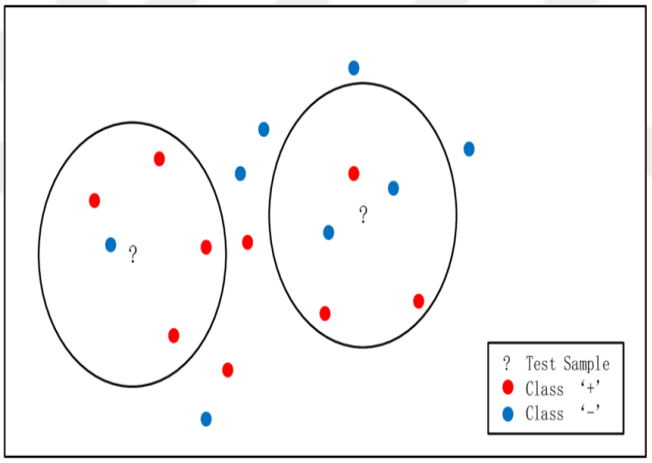

Figure 3.8. A representation of K-NN classification (de Carvalho Couto et al., 2022)

3.7.4. Classification using a DT

The decision-Tree algorithm is extensively utilized ML algorithm regarding issues of both classification and regression. Mainly, the algorithm replicates similar to humans' thinking, making it quite prevalent and simpler to construe the inputs with better consideration of the issue (Olanow et al., 2001). In this algorithm, a DT represents a tree among its node denotes to the characteristic while its association states to a decision rule along with its leaf node states to an output category. The purpose is to yield a tree-like assembly for the input features and build a distinctive output at each leaf (Pandya & Pandya, 2015).

Furthermore, the requirement of implementing the decision-tree algorithm in our work is to create a training model that is supposed to be applied for the estimate of output category

by supposing the decision rules are understood from the primary trained data. Figure 3.9 represents the structure of the algorithm.

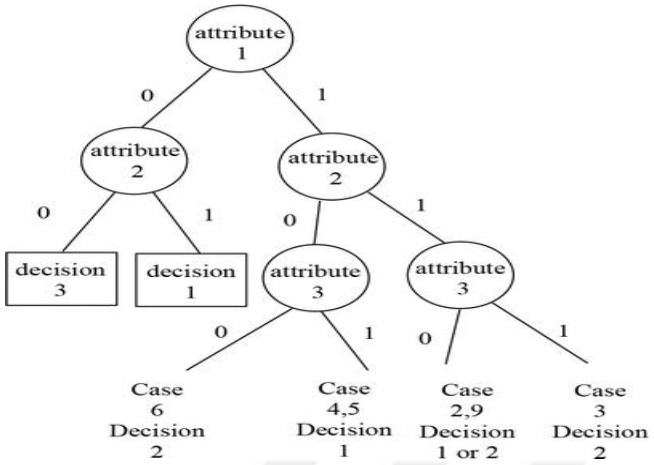

Figure 3.9. DT Representation, (Kutschenreiter-Praszkiewicz, 2017)

3.7.5. Classification using RF

A RF illustrated in Figure 3.10 can be described as a ML technique that may be used to solve problems such as regression and classification (Breiman, 2001). This classifier is extensively employed in various areas with motivating performances, predominantly in medical imaging (Parmar et al., 2015).

RF takes advantage of ensemble methods, i.e., combining various classifiers. An RF method is encompassed numerous DTs. The 'forest' of the RF method is taught using bagging. Bagging can be described as the meta-algorithm-based ensemble that enhances the results of ML methods by increasing the number of algorithms in the ensemble. Because the DT provides predictions, the RF algorithm may decide the outcome based on those predictions. It anticipates by taking an average of the output of numerous trees simultaneously. Using a RF method alleviates the limits of the DT technique. It reduces the amount of overfitting in the dataset and increases accuracy.

The rule is to create a multitude (k) of independent trees that are made from an initial sample that conforms to N patients with F features. The preliminary training sample might be characterized through a matrix of *size* (N, F). After building, the forest is then applied for classification making predictable labels such as label 0 as respondent and label 1 as non-respondent. Subsequently, the final labeling of patients is performed with most of the votes from the forest. Two arbitrary procedures are employed for constructing the forest. In the first process, every tree of the forest is created out of a bootstrap sample of k patients, randomly selected with a substitute. Hence that a bootstrap might have numerous times similar patient info. In the second process, out of this sample, a binary tree is built out of the DT. For every tree node, a subclass of f features selected arbitrarily between the F features is outlined. Here

f is equal to the rounded $\sqrt{F}$ (Genuer et al., 2010). Moreover, the cross-validation technique is utilized for assessing the classifier according to a sampling method, signifying that the dataset is split into K subsamples. $K - 1$ are applied to create training samples, whereas the last is deemed the test sample. This process is comprehended K times with variation to employ each subsample as an authentication set. Through comparing estimated labels and truth, capable of estimating accuracy for every test sample. The model accuracy resembles the standard deviation (SD) and the mean of K computed accuracies.

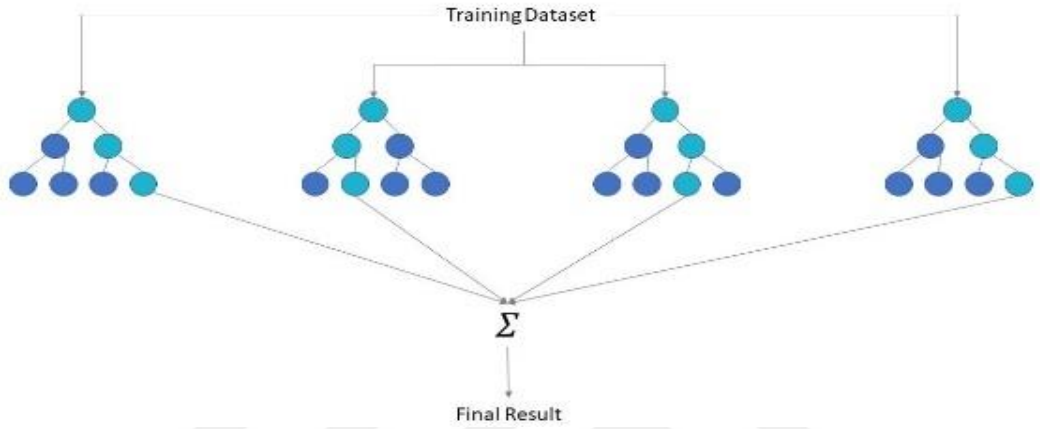

Figure 3.10. RF depiction, (Ghorbanzadeh et al., 2020)

3.7.6. Classification using XGBoost

The initial development of XGBoost began as a research effort by Tianqi Chen (Chen & Guestrin, 2016), who was working as a member of the Distributed ML Community. In its early stages, it was only a terminal program that could be customized via the LIBSVM configuration file (Chih & Chih, 2022). Afterward, it received massive recognition in the ML competition community. This increased the number of developers who could use the library and increased its popularity in several tournaments.

XGBoost, illustrated in Figure 3.11, has outstanding scalability and high running speed, making this algorithm an effective machine-learning technique (Chen & Guestrin, 2016). In this work, a tree booster is utilized for each iteration. Moreover, this classifier is prevalent since it is consistent, effectual, predictable, and moderately slow in execution (Chen et al., 2015). This course familiarizes me with numerous of Kaggle's challenges.

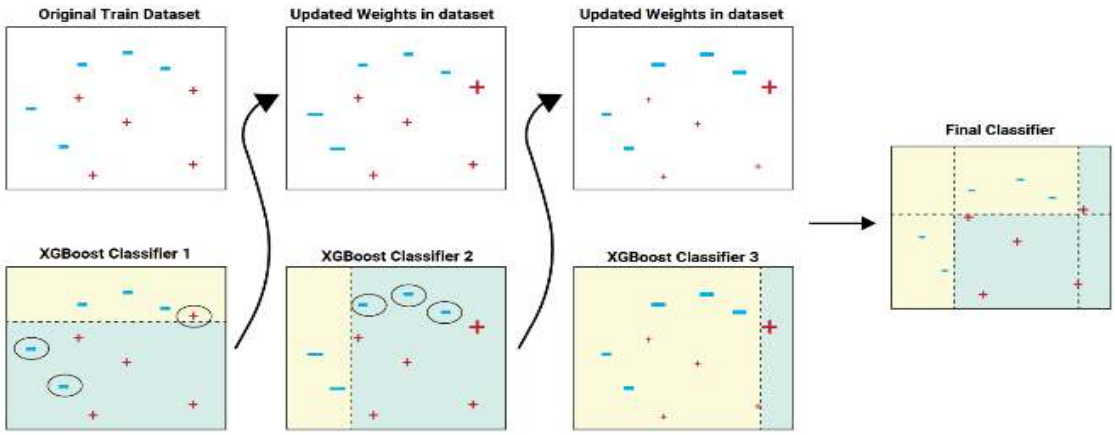

Figure 3.11 XG-Boost classification example (Ishan, Shah, 2020)

3.7.7. Classification using Gaussian Naive Bayes (Gaussian-NB)

Gaussian Naive Bayes is a variation of Naive Bayes that uses the Gaussian-based normal distribution and works with continuous numbers. It is a basic categorization approach with much power. When the dimensionality of the inputs is high, they are helpful. The Naive Bayes Classifier may also be used to solve complex classification. The classification is performed through the Bayes principle to measure the probability of class name C, given that the case $X_1 \dots X_n$ by the formula $\mathrm{P}(C = c \mid X_1 = X_1, \dots\dots, X_n = x_n)$. The classifier can be determined as Classify $(\mathrm{B}_1 \dots \mathrm{B}_n) = \operatorname{argmax} \mathrm{P}(\mathrm{C} = \mathrm{c}) \prod_{i=1}^{n} p(B_i = bi \mid C = c)$ where $(B_1 \dots B_n)$= Attribute Variables and C = Class name (M. & P., 2016). Furthermore, training data are divided through class and mean, and the standard deviation of each class is measured. So, for calculating the probabilities of the dataset following equation can be applied

$$\mathrm{P}(X = x \mid C = c) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{3.6}$$

Where x= variable, c= class, σ= standard deviation.

3.8. Classification using DNN

DL, a form of ML, is built on NN, which employs computer approaches that replicate the workings of the human brain. NNs use training to comprehend the underlying structure of the data they are given. Then the recognized patterns are used to produce predictions for fresh data sets like the original. In this second wording, a NN depicted in Figure 3.12 is described as the functional unit of DL. It is designed to replicate the human brain and recreate its mental processes to tackle complicated data-driven challenges (Ngiam et al., 2011).

Furthermore, NN has extensive uses in numerous areas, predicting medical applications (Azar, 2013). Previously NN has been utilized for cancer classification (J. Khan et al., 2001) and areas of bioinformatics [19–21]. Additional utilities of ANN models have been discussed in [9–11, 22–24] and in citations.

In Figure, the hidden layer processes the inputs through a connection function described as follows,

$$h_{W,b}(X) = f(W^T X + b) \tag{3.7}$$

Where W refers to weight and b to bias, all input layer neurons are active. In that case, every input neuron will increase its corresponding weight matrix, and the output will be added with a bias, then given to a nearby hidden layer.

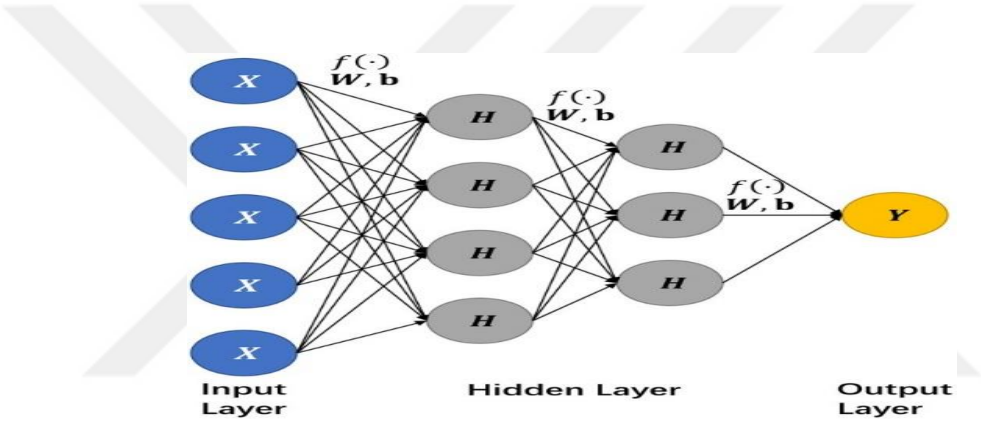

Figure 3.12. Neural Network structure (Tang et al., 2019)

Explicitly, the input of the activation function is a blend $(W^T X + b)$ Symbolized in equation (3.7), function output is then given to an adjacent neuron as a new input. In the connection formula, the previous input feature might be removed to the next layer; through this way, the components might be properly extracted and refined. The overall feature extraction performance relies considerably on the preference of the activation function.

3.9. Classification using Deep-RL

Deep-RL (Puzanov et al., 2020) is a branch of ML and AI that allows intelligent computers to learn from their actions, like through experience. Incentives are given for actions that lead to the desired outcome reinforced. A machine continues to learn via trial and error, which makes this technique ideal for complex systems that are constantly evolving; At the same time, RL has been known for decades; it has just recently been coupled with DL, with stunning results. The "deep" component of RL refers to the layers of artificial neural networks that are "deep" in comparison to the human brain's structure (Francois-Lavet et al., 2018). DL requires

massive amounts of training data and considerable computer resources. The volume of data has risen in recent years, while the cost of processing resources has decreased dramatically, enabling DL applications to emerge (Liu et al. 2019). Clinical trials such as cancer dedication and categorization, novel medication research and automated therapy, and Deep-RL promise to enhance healthcare (Cohn et al., 2018). Figure 3.13 depicts Deep-RL.

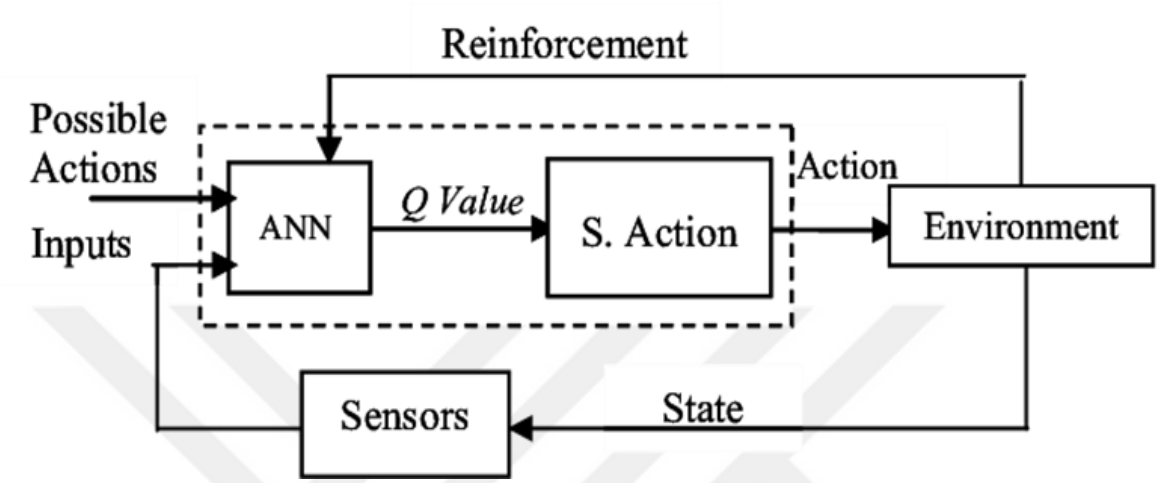

Figure 3.13. Deep-RL (Hatem & Abdessemed, 2009) (Foudil Abdessemed 2009)

3.10. Performance Assessments

In ML applications, recall and other specificity measures are extensively utilized besides the common accuracy measure—formal definitions for these performance criteria are labeled as shown below.

- True Positive (*TP*) of positive samples correctly predicted to be positive.
- True Negative (*TN*) of negative samples correctly predicted to be negative.
- False Positive (*FP*) of negative samples erroneously predicted to be positive.
- False Negative (*FN*) of positive samples erroneously predicted to be negative.
- (*N*) number of sample instances.

3.11. Confusion Matrix

The confusion matrix is utilized for evaluating the classification model. This is fundamentally an effective classification model that is proficient in summarizing the model's efficiency at categorizing examples into diverse classes. Table 3.1 shows the confusion matrix representation.

		Predicted	
		P	**N**
Actual	P	True Positive	False Negative
	N	False Positive	True Negative

Figure 3.14 Confusion matrix

- **Accuracy**: It is a measure of the ratio of correctly predicted.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{3.1}$$

- **Precision**: The ratio of the true positive number of total numbers of forecasted positive samples. Precision is TP / predicted positive.

$$Precision = \frac{TP}{(TP + FP)} \tag{3.2}$$

- **Recall or sensitivity**: The ratio of positive samples was correctly stated in the total number of positive samples, known as sensitivity analysis of the true positive ratio (TPR).

$$Recall = \frac{TP}{(TP + FN)} \tag{3.3}$$

- **F-Measure**: It is the average accuracy to recall. It is an important measure that combines precision and recall.

$$F - score = \frac{(2 * Recall * Precision)}{(Recall + Precision)} \tag{3.4}$$

Receiver Operating Characteristic (ROC): Active part of approximating the performance of diagnostic tests. ROC is the relationship between specificity and sensitivity for a specific diagnosis. For the ROC curves, the y-axis denoted sensitivity, and the x-axis represented specificity. TPR is plotted over the y-axis, and FP is plotted along the x-axis.

Furthermore, the ROC curve provides a graphical diagram of trade-offs between sensitivity (TRP) and specificity (1- FPR). The Classifiers that imply more robust results lead to curves closer to the top-left corner.

3.12. Tools

This thesis utilized Google Collaboratory, a cloud-based service based on Jupiter notebooks, to implement our task. Collab provides 12 GB of RAM and can be increased up to 25 GB. All the experiments are performed with Tensorflow and Keras framework on NVIDIA Tesla K80 GPU using Google Collab.

4. EXPERIMENTAL RESULTS

4.1. Framework

This thesis aims to outline possibilities to explore disease through ML algorithms on datasets in medical care. The dataset comprises clinical data taken from both diseased and healthy tissues. The first step for applying the ML techniques is to prepare the cancer dataset, and the next step states which algorithm perfume better.

This thesis work utilized supervised ML techniques and Deep-RL to improve the process of classifying cancer genes based on microarray gene expression. Associating both methods to improve cancer classifying personalized to a particular cancer genome. The gene expression dataset aids in recognizing new cancer cases and provides a more standard approach to help find new classes of cancer and tumor classification. The data was further analyzed to determine the classification of ALL and AML.

4.2. Performance Characteristics

Aiming to evaluate the performance of utilized models for classifying the dataset, it's logical to integrate performance characteristics, for example, classification accuracy or true/false positive rate, used for evaluating and comparing the outcomes obtained throughout runs by different models. The performance measures of the ROC are presented to address additional characteristics of the results achieved.

This thesis utilized eight supervised ML models and compared the performance of each model using three different feature selection techniques and compared them with the non-PCA (raw dataset). Since PCA permits a selection of primary components appearing in converted outcomes (Abdi & Williams, 2010). PCA has recently been utilized in many healthcare-related studies (Reddy et al., 2020; Shakil et al., 2020). Also, PCA is utilized to regulate the comparative relevance of features. Further, we also tested our models without utilizing PCA and compared the results. By utilizing the PCA pre-processing technique, we drastically reduced the dataset's dimensionality. Additionally, six different PCA Components are applied (PCA 2, 10, 14, 20, 30, 38), but only the PCA 2 component is utilized in the results because it achieved better results than the rest.

Nevertheless, Lasso regularization is an embedding feature selection technique based on a linear regression model. Its high efficacy has fascinated widespread consideration in the feature selection field. Lasso decreases the original feature space's dimensions by compressing and selecting variables. When creating a linear regression model, under the limitation that the sum of absolute values of the regression coefficients is much less than an actual threshold, then the regression coefficients with less absolute values are automatically compressed to 0 value

(Wang et al., 2022). After these feature variables are removed, the purpose of reducing the dimension of feature space is accomplished.

Moreover, in this trained network thesis with 50 epochs, the training loss is 1.95%. Further, an early stopping mechanism was employed to stop training on the 12th epoch since the training accuracy reached 100%; however, training loss was still high and overfitted.

4.3. Results of ML-based Classifiers

4.3.1. Results of LR Classifier

We have tried Hyperparameter optimization for LR using scaled data with grid search values of C-index from $[1e-03, 1e-2, 1e-1, 1, 10]$, with penalty values of 11 and 12. The model achieved 91.17% accuracy, and the non-PCA version of the LR model achieved 91.17 accuracies.

Further, to improve the process of classifying cancer genes based upon microarray gene expression, the achieved by utilized models, with the scaled data along with grid search to help fine-tune the model. LR obtained similar results with and without PCA. The classification results from the LR classifier are tabulated in Table 4.1. Figure 4.1 shows the confusion metric of the logistics regression and ROC curve. A confusion matrix, also known as an error matrix, is a specific table that gives a visualization of performing an algorithm, typically supervised learning. The classification results remain the same, concluding that there is no effect on the results if we opted for PCA or used PCA.

Besides, the RF importance version also did not perform well enough, which has achieved lower accuracy than the non-PCA and PCA model versions. Lastly, the Lasso regularization-based model overfitted every expectation and performed worst. Figure 4.2 shows the confusion metric and ROC curve of the RF importance-based LR model, and 4.3 shows the confusion metric and ROC curve of the lasso regularization-based LR model.

Table 4.1. Classification report of LR model

Methods	Accuracy %	ROC Value%	Precision (Avg.)%	Recall (Avg.)%	F1-score(Avg.)%
LR (non-PCA)	97.00	97.05	96.05	97.05	97.00
LR (PCA)	91.17	91.42	91.00	91.50	91.00
LR (RF Importance)	85.00	86.42	85.00	86.00	85.00
LR (Lasso Regularization)	100	100	100	100	100

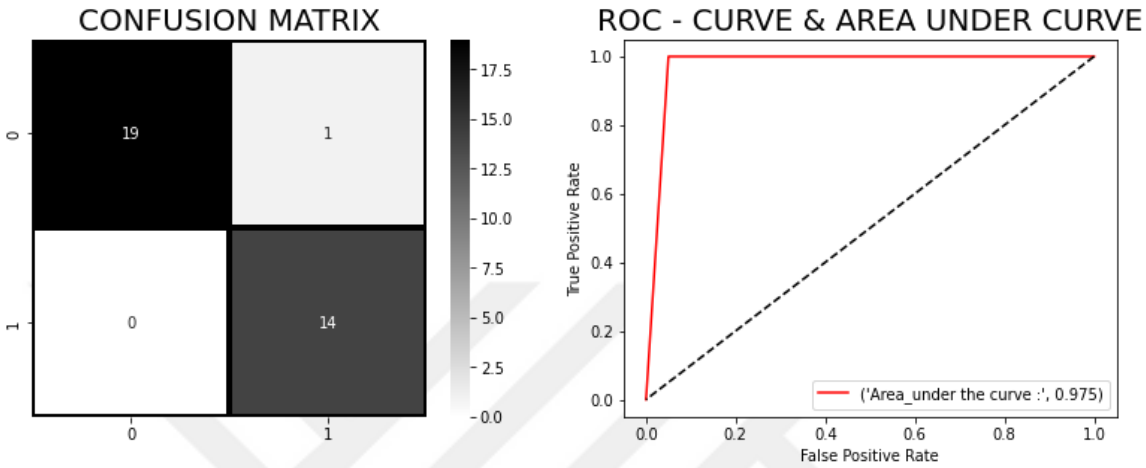

Figure 4.1. Confusion metric and ROC curve for LR model (Identicaly results in both PCA and non-PCA versions)

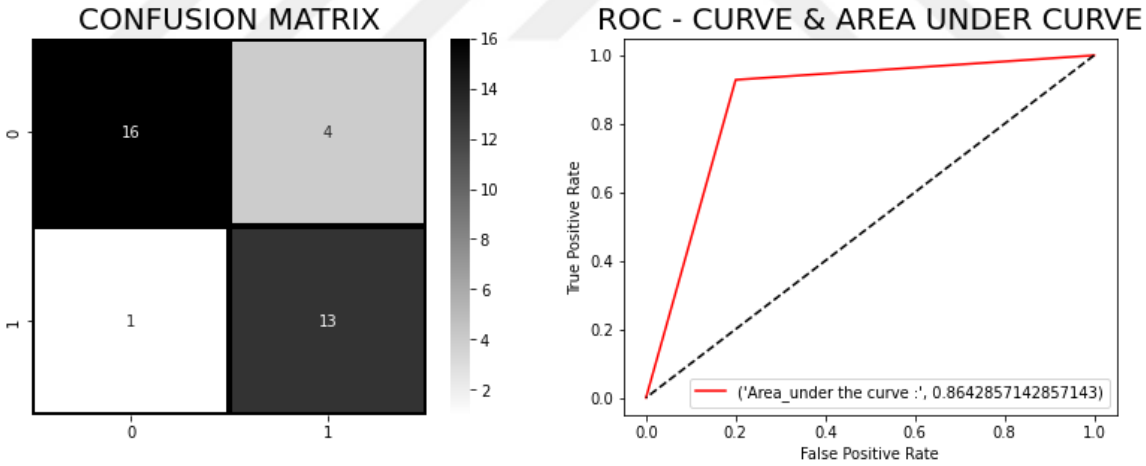

Figure 4.2. Confusion metric and ROC curve of LR model (RF importance)

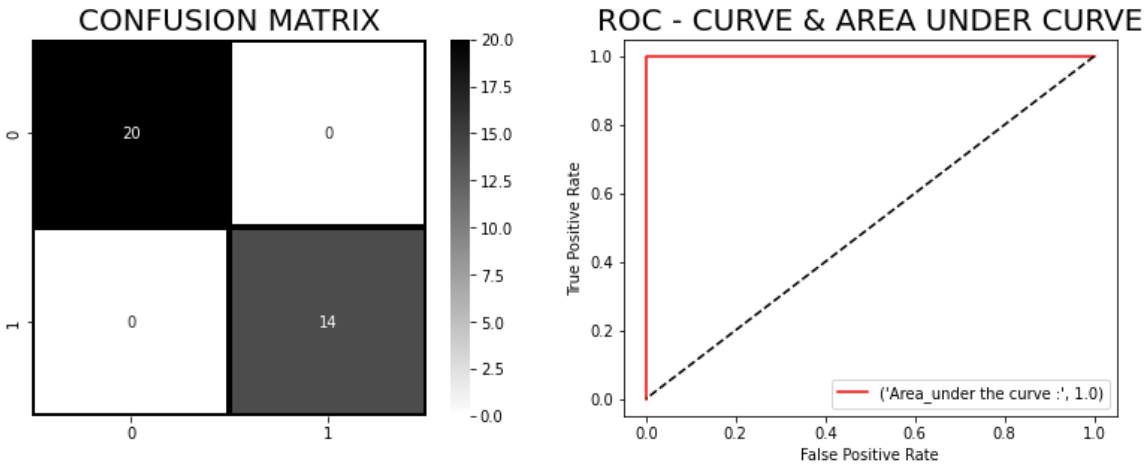

Figure 4.3. Confusion metric and ROC curve of LR model (lasso regularization)

4.3.2. Results of SVM Classifier

Afterward, another classifier, the SVM classifier. The SVM also used the PCA version of the dataset along with the original dataset without PCA. Again, like LR, grid search cross-validation was used to tune the model.

$svm\_param\_grid = \{'C': [0.1, 1, 10, 100]$, $gamma = [1, 0.1, 0.01, 0.001, 0.00001, 10]$, $kernel = ["linear", "rbf", "poly"]$, $decision\_function\_shape = ["ovo", "ovr"]$. Best Parameters in this model $\{'C': 0.1, 'decision\_function\_shape': 'ovo', 'gamma': 1, 'kernel': 'linear'\}$.

SVM provided an accuracy of 91.20% with a ROC value of 91.40 with PCA utilization. But with the non-PCA, the accuracy of SVM improved to 97.00% with the same ROC curve value of 99.46 illustrated in Figure 4.4. Furthermore, the classification results from the SVM classifier are tabulated in Table 4.2. Figure 4.5 shows the confusion metric of the SVM and ROC curve for the PCA version.

Furthermore, the RF importance version achieved an accuracy of 97% and a ROC value of 97.50, which is on par with the non-PCA model. On the other hand, this configuration achieved better accuracy than the PCA-based model. Figure 4.6 shows the confusion metric of the SVM and ROC curve for the RF importance-based model. However, the lasso regularization version is overfitted, similar to the LR model. Figure 4.7 confusion metric of the SVM and ROC curve for the lasso regularization-based model.

Table 4.2. Classification report of the SVM model

Methods	Accuracy %	ROC Value%	Precision (Avg.)%	Recall (Avg.)%	F1-score (Avg.)%
SVM (non-PCA)	94.12	99.46	0.97	0.97	0.97
SVM (PCA)	91.20	91.40	91.00	91.50	91.00
SVM (RF Importance)	97.00	97.50	97.00	97.00	97.00
SVM (Lasso Regularization)	100	100	100	100	100

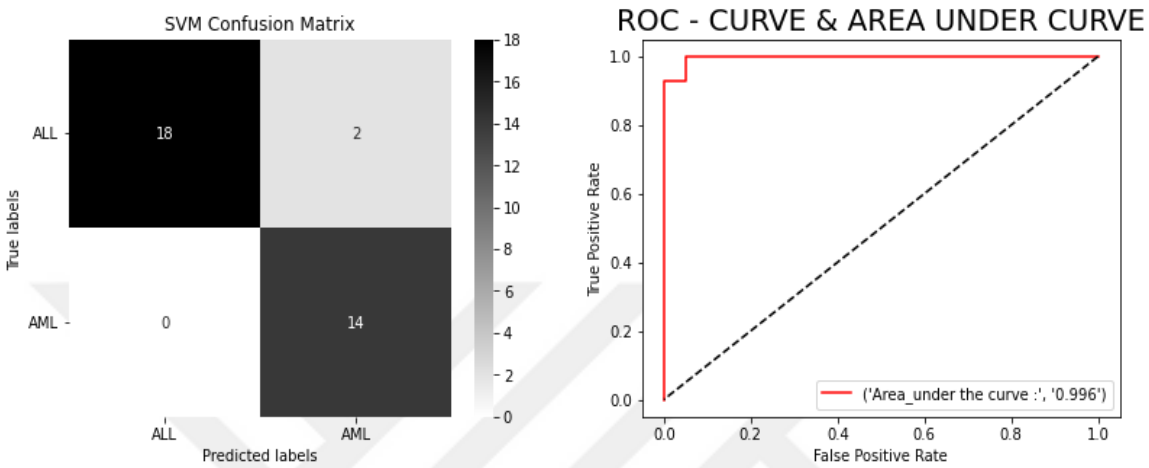

Figure 4.4. Confusion Metric and ROC curve for SVM model (non-PCA)

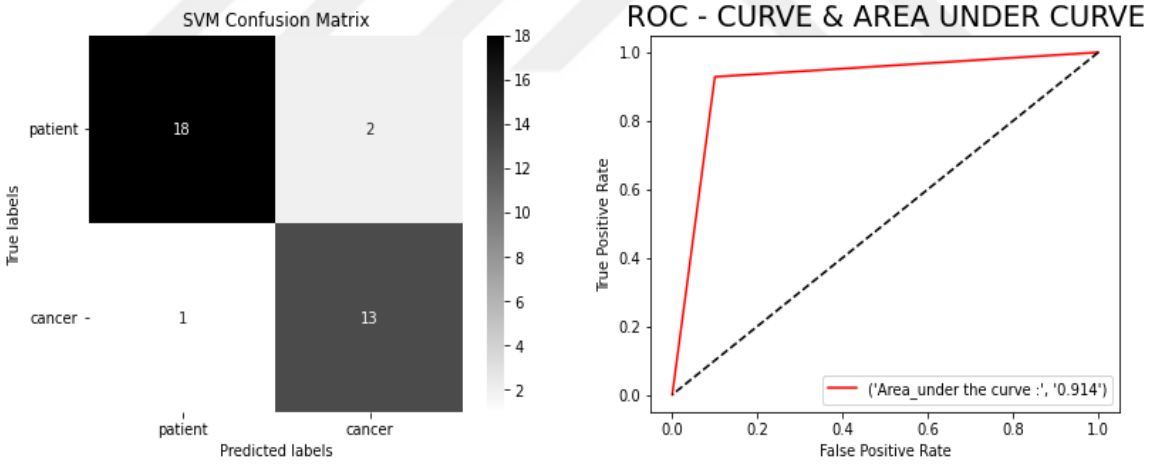

Figure 4.5. Confusion Metric and ROC curve for SVM model (PCA)

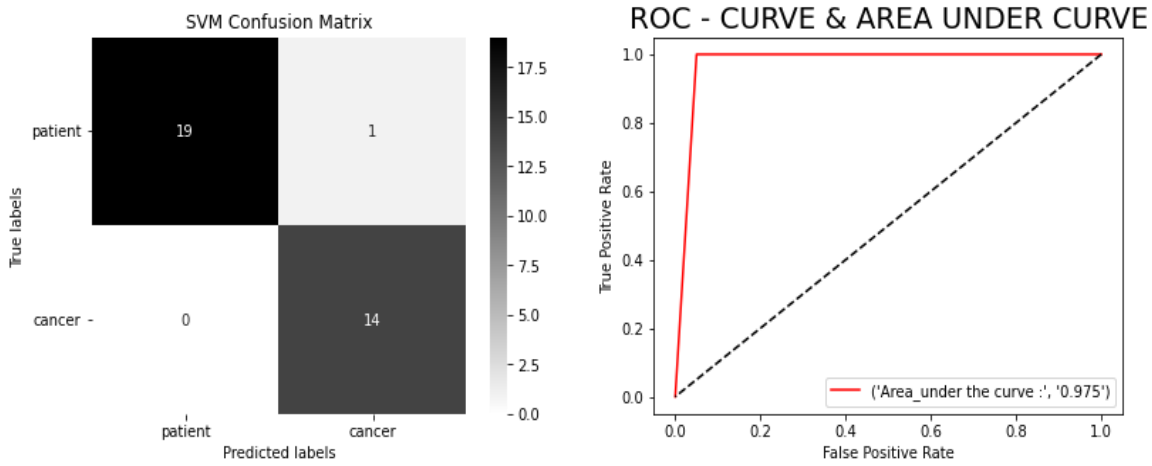

Figure 4.6. Confusion metric and ROC curve of SVM model (RF importance)

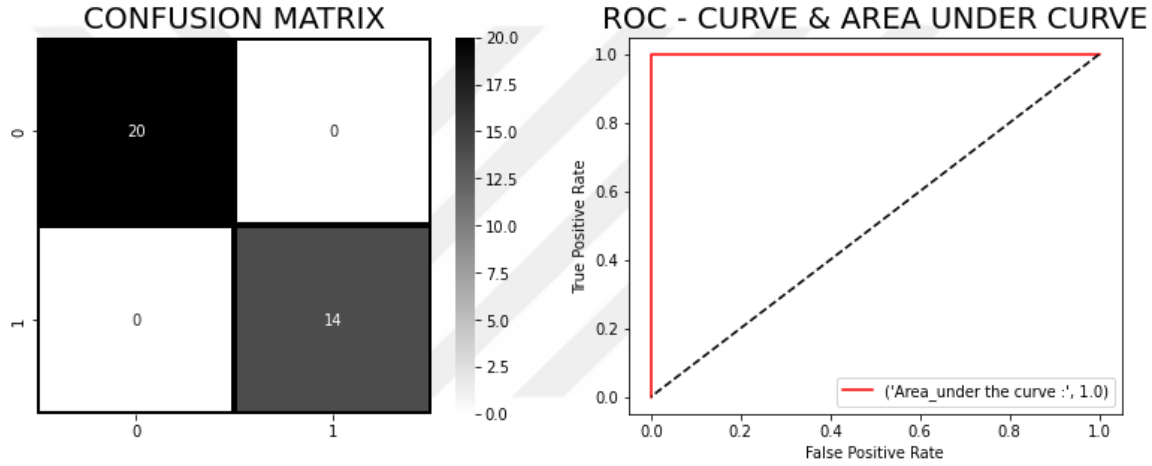

Figure 4.7. Confusion metric and ROC curve of SVM model (lasso regularization)

4.3.3. Results of K-NN Classifier

The next classifier is K-NN with $n\_neighbors = [i\ for\ i\ \boldsymbol{in}\ range\ (1,30,5)]$, with uniform weights, algorithms from ["$ball\_tree$","kd_tree","$brute$"] the leaf from $[1, 10, 30]$, and the $p = [1,2]$. The non-PCA of the K-NN classifier achieved an accuracy of 82.35% with a ROC value of 80.71. on the contrary, the PCA version obtained an accuracy of 85.29%. Besides, the classification results on both PCA and non-PCA obtained from the K-NN classifier are tabulated in Table 4.3. Figure 4.8 shows the confusion metric of the non-PCA version of K-NN and ROC curve values, whereas Figure 4.9 shows the PCA version, respectively. As a result, the PCA version has achieved higher accuracy and ROC value than the non-PCA.

Furthermore, the dataset based on RF importance achieved an accuracy of 94% with a ROC value of 93.92, which is higher than the non-PCA and PCA versions. Figure 4.10 shows the confusion metric of K-NN and ROC curve values for the RF importance-based model. Besides, the Lasso Regularization-based K-NN model achieved an accuracy of 91%, which is

also higher than the non-PCA and PCA versions of the K-NN model. In this configuration, the RF importance-based dataset utilized in the K-NN model performed better in every aspect. Figure 4.11 shows the confusion metric of K-NN and ROC curve values for the lasso regularization-based model.

Table 4.3. Classification report of the K-NN model

Methods	Accuracy %	ROC Value%	Precision (Avg.)%	Recall (Avg.)%	F1-score (Avg.)%
K-NN (non-PCA)	82.35	80.71	83.00	81.00	81.00
K-NN (PCA)	85.29	85.35	84.50	85.50	85.00
K-NN (RF Importance)	94.00	93.92	94.00	94.00	94.00
K-NN (Lasso Regularization)	91.00	89.28	93.00	89.00	91.00

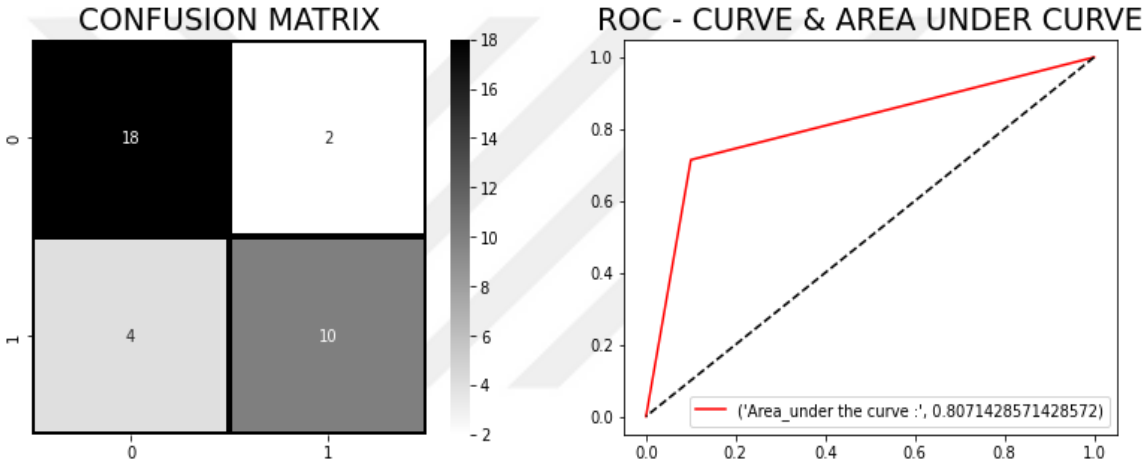

Figure 4.8. Confusion Metric and ROC curve for K-NN model (non-PCA)

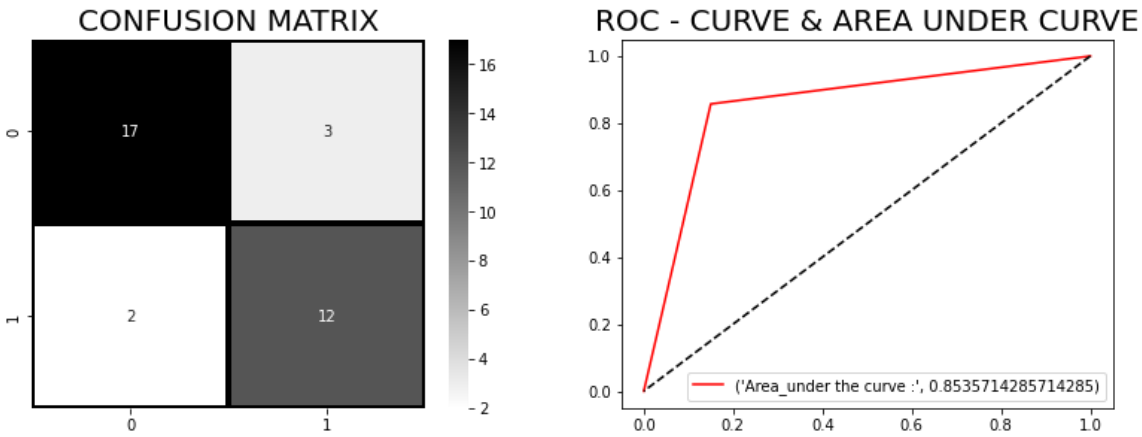

Figure 4.9. Confusion Metric and ROC curve for K-NN model (PCA)

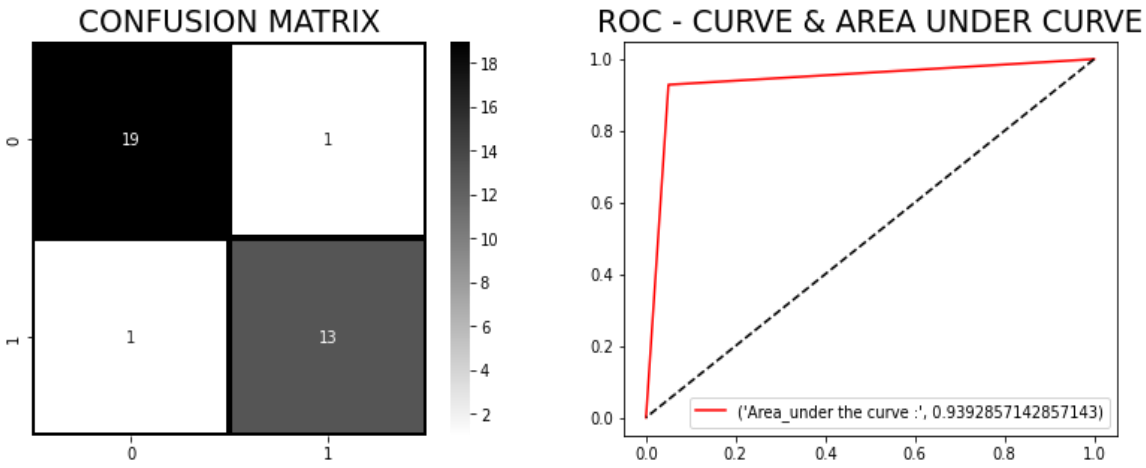

Figure 4.10. Confusion metric and ROC curve of K-NN model (RF importance)

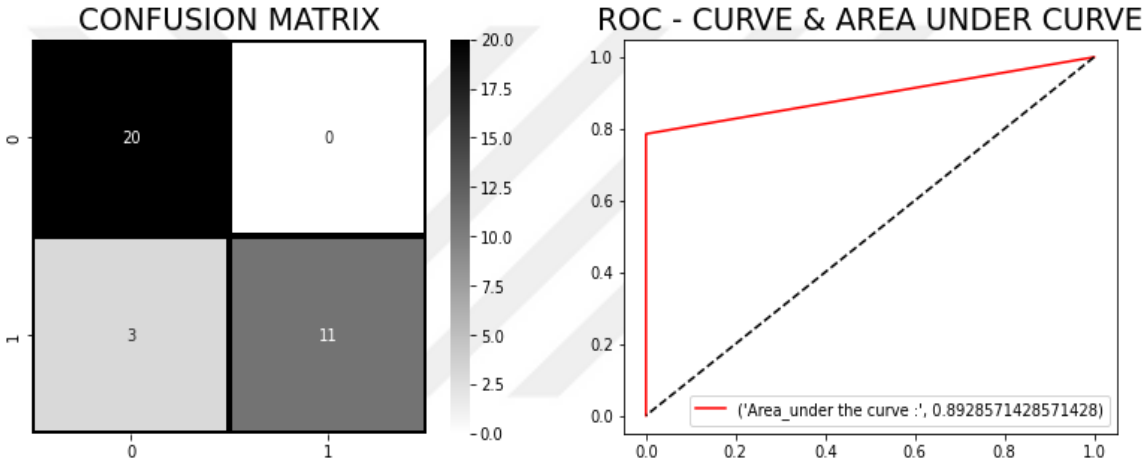

Figure 4.11. Confusion metric and ROC curve of K-NN model (lasso regularization)

4.3.4. Results of DT Classifier

In this thesis, another classified applied, which is a DT with the following parameters. $params$ = {'max_$leaf_nodes$': $list$ ($range$ (2, 100)), 'min_$samples_split$': [2, 3, 4, 5, 6], 'max_depth': [,4,5,6,7,8]. The DT model with non-PCA configuration obtained an accuracy of 91.18% with a ROC value of 91.42%. However, the PCA version of the DT model obtained 88.23% with a ROC value of 88.92; Table 4.6 shows the classification report for both non-PCA and PCA versions of the DT classifier. As we can see, the non-PCA version of the model performed better. The confusion matrix and ROC curve value of the non-PCA model are illustrated in Figure 4.12. The ROC value and confusion matrix is shown in Figure 4.13 for the PCA version.

Additionally, based on RF importance, the DT model achieved an accuracy of 91% with a ROC value of 91.42, which is higher than the PCA version but achieved similar results with non-PCA, a raw dataset model configuration. Figure 4.14 shows the confusion metric of

the DT model and ROC curve values for the RF importance-based model. Besides, the Lasso Regularization-based DT model achieved an accuracy of 91%, which is also higher than the PCA version of the DT model but achieved similar results as the non-PCA and RF importance-based models, respectively. In this configuration, the RF importance-based dataset, non-PCA version (raw dataset), and lasso regularization utilized in the DT model performed better except for the PCA version of the model. Figure 4.15 shows the confusion metric of the DT model and ROC curve values for the lasso regularization-based model.

Table 4.4. Classification report of DT model

Methods	Accuracy %	ROC Value (Avg.)%	Precision (Avg.)%	Recall (Avg.)%	F1-score (Avg.)%
DT (non-PCA)	91.18	91.42	91.00	91.00	91.00
DT (PCA)	88.23	88.92	87.50	89.00	88.00
DT (RF Importance)	91.00	91.42	91.00	91.00	91.00
DT (Lasso Regularization)	91.00	91.42	91.00	91.00	91.00

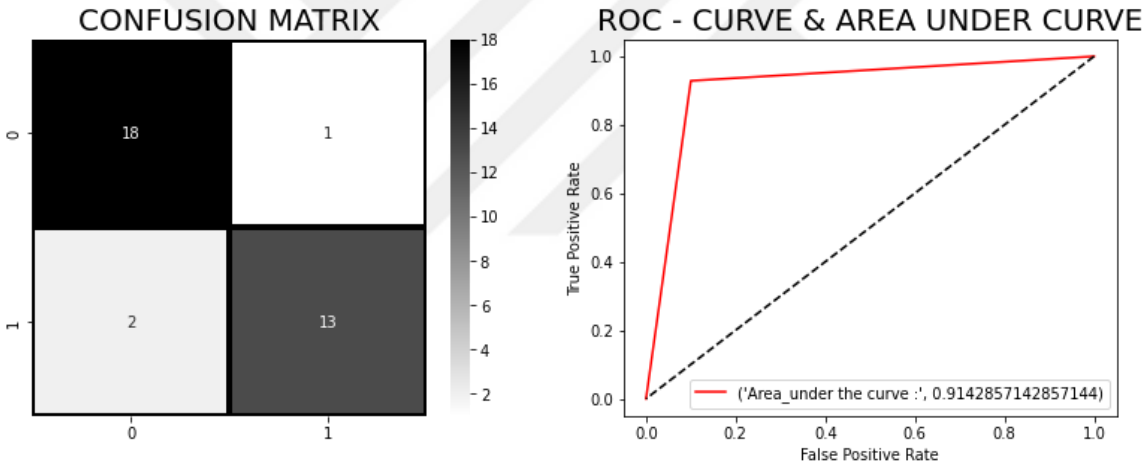

Figure 4.12. Confusion Metric and ROC curve for DT model (non-PCA)

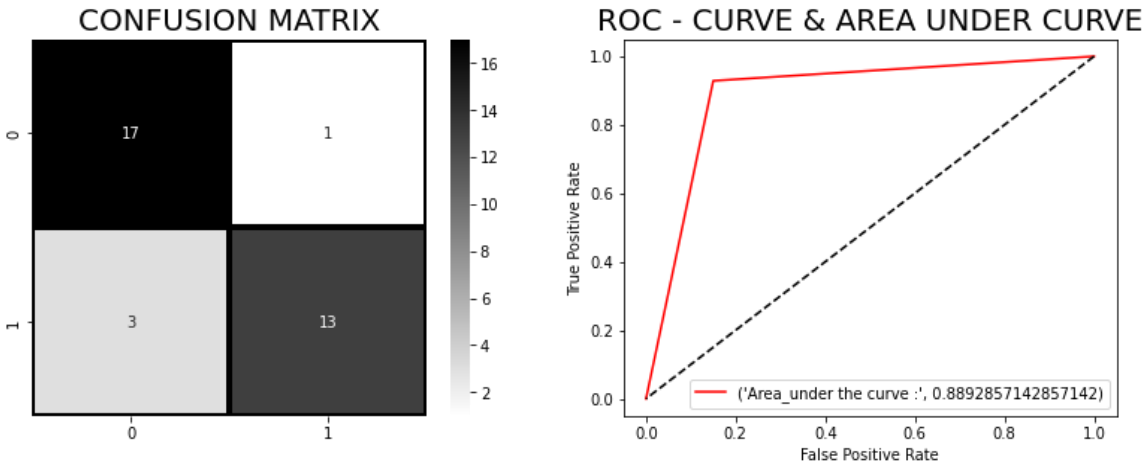

Figure 4.13. Confusion Metric and ROC curve for DT model (PCA)

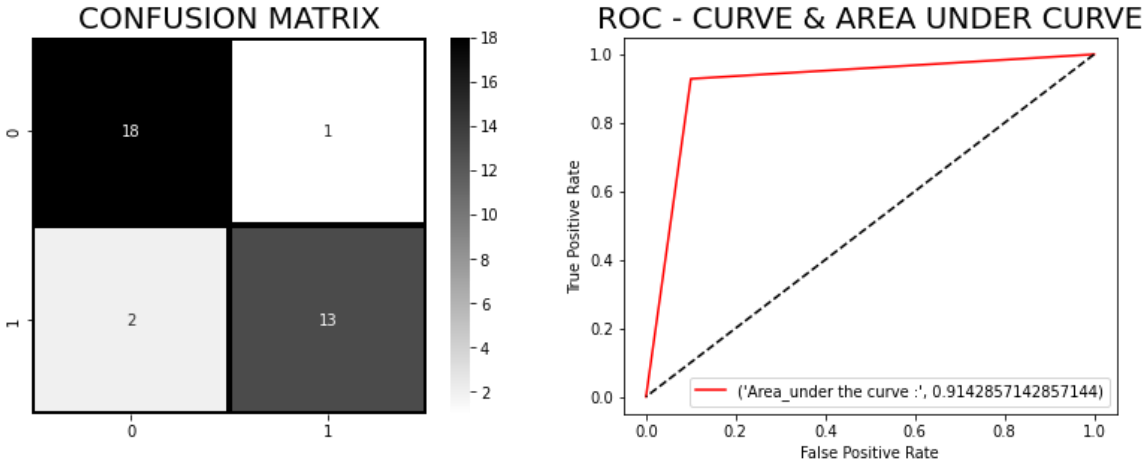

Figure 4.14. Confusion metric and ROC curve of DT model (RF importance)

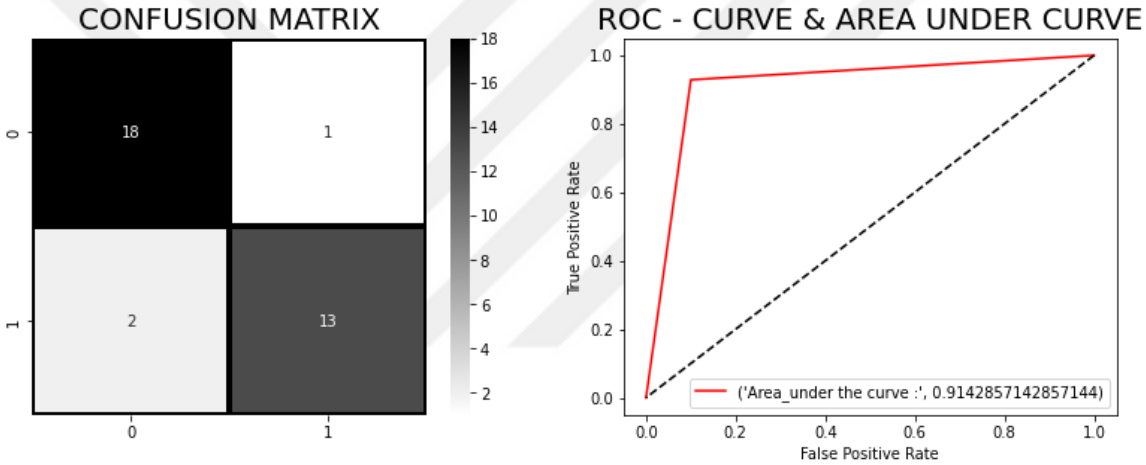

Figure 4.15. Confusion metric and ROC curve of DT model (lasso regularization)

4.3.5. Results of RF Classifier

To evaluate the dataset, employed RF classifier with $n\_estimators = [60, 70, 80, 90, 100]$ $max\_features = [0.6, 0.65, 0.7, 0.75, 0.8]$ with $min\_samples\_leaf = [8, 10, 12, 14]$ along with $min\_samples\_split = [3, 5, 7]$. The best RF was with the $max\_features = 0.6, min\_samples\_leaf = 8, min\_samples\_split = 3, and\ n\_estimators = 70$. Again, used grid search cross-validation to tune the model like previous models. Like other models, the non-PCA obtained an accuracy of 91.17% with the ROC value of 91.42, whereas the PCA version of RF obtained decreased accuracy of 88.23%. The classification results obtained from the RF classifier are tabulated in Table 4.5. RF model also results like the DT model where non-PCA (raw dataset) performed better than the PCA utilized formation of the model. Figure 4.16 shows the confusion metric of the RF model along with the ROC curve value for non-PCA, and Figure 4.17 illustrates the ROC value and confusion matrix for the PCA version of the model.

Also, based on RF importance, the RF model achieved an accuracy of 91% with ROC value of 91.42, which is higher than the PCA version but achieved similar results with non-PCA model configuration. Figure 4.18 shows the confusion metric of the DT model and ROC curve values for the RF importance-based model. Besides, the Lasso Regularization-based RF model achieved an accuracy of 91%, which is also higher than the PCA version of the RF model but achieved similar results as non-PCA and RF importance-based models, respectively. In this configuration, the RF importance-based dataset, non-PCA version (raw dataset), and lasso regularization utilized in the RF model performed better except for the PCA version of the model. Figure 4.19 shows the confusion metric of RF and ROC curve values for the lasso regularization-based model.

Table 4.5. Classification report of RF model

Methods	Accuracy %	ROC Value (Avg.)%	Precision (Avg.)%	Recall (Avg.)%	F1-score (Avg.)%
RF (non-PCA)	91.17	91.42	91.00	91.00	91.00
RF (PCA)	88.23	87.85	90.00	90.00	90.00
RF (RF Importance)	91.00	91.42	91.00	91.00	91.00
RF (Lasso Regularization)	91.00	91.42	91.00	91.00	91.00

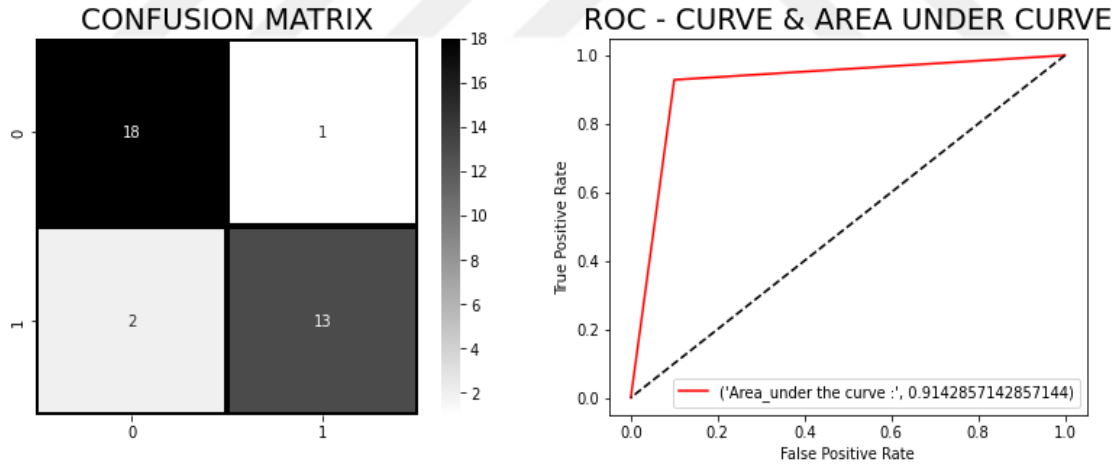

Figure 4.16. Confusion Metric and ROC curve for RF model (non-PCA)

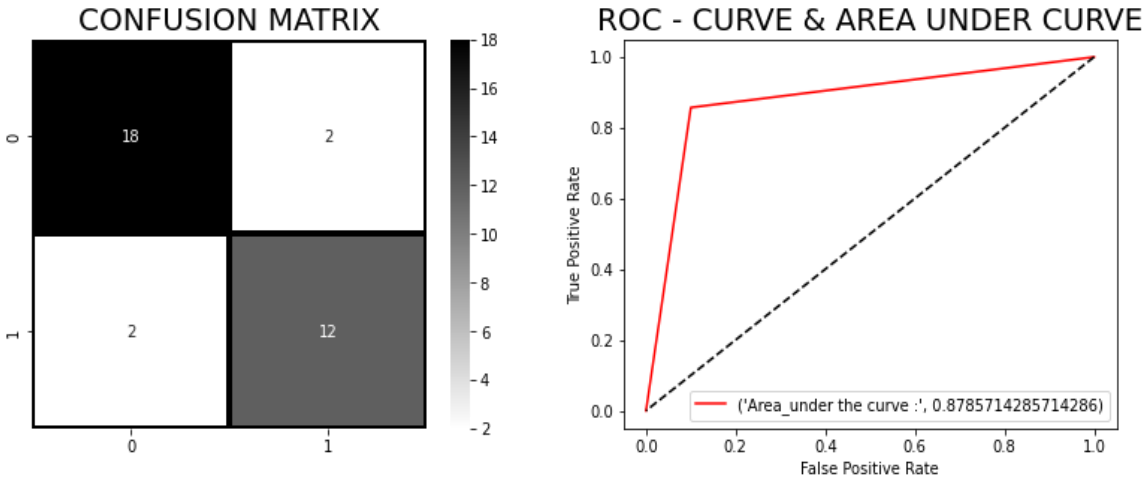

Figure 4.17. Confusion Metric and ROC curve for RF model (PCA)

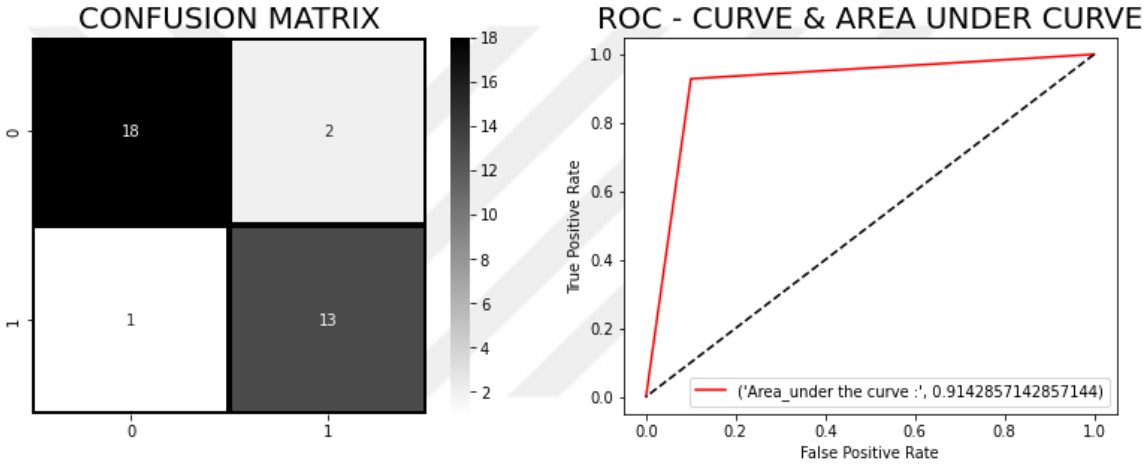

Figure 4.18. Confusion metric and ROC curve of RF model (RF importance)

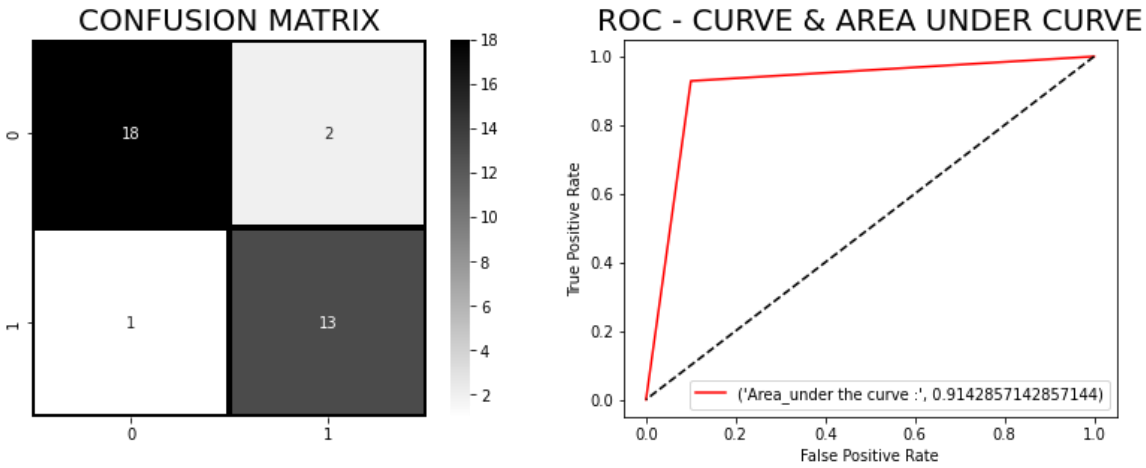

Figure 4.19. Confusion metric and ROC curve of RF model (lasso regularization)

4.3.6. Results of XG-Boost Classifier

Table 4.6 shows classification results obtained from the XG-Boost classifier for the non-PCA and PCA model versions. The validation accuracy of the non-PCA version is 91.17% which is identical to both DT and RF models, along with ROC curve value of 91.42. Figure 4.20 shows the confusion metric and ROC curve value for the non-PCA version of the XG-Boost model.

Moreover, the XG-Boost model is also seen as decreased performance with PCA utilization, which is similar to the previous two models: DT and RF. Figure 4.21 shows the ROC and confusion matrix.

Similarly, based on RF importance, the XG-Boost model achieved an accuracy of 91% with ROC value of 91.42, which is higher than the PCA version but achieved similar results with non-PCA model configuration and lasso regularization. Figure 4.22 shows the confusion metric of the XG-Boost model and ROC curve values for the RF importance-based model. Besides, the Lasso Regularization-based XG-Boost model achieved an accuracy of 91%, which is also higher than the PCA version of the XG-Boost model but achieved similar results as of non-PCA and RF importance-based models, respectively. In this configuration, the RF importance-based dataset, non-PCA version (raw dataset), and lasso regularization utilized in the XG-Boost model performed better except for the PCA model version. Figure 4.23 shows the confusion metric of XG-Boost and ROC curve values for the lasso regularization-based XG-Boost model.

Table 4.6. Classification report of XG-Boost model

Methods	Accuracy %	ROC Value (Avg.)%	Precision (Avg.)%	Recall (Avg.)%	F1-score (Avg.)%
XG-Boost (non-PCA)	91.17	91.42	91.00	91.00	91.00
XG-Boost (PCA)	88.24	87.85	90.00	90.00	90.00
XG-Boost (RF Importance)	91.00	91.42	91.00	91.00	91.00
XG-Boost (Lasso Regularization)	91.00	91.42	91.00	91.00	91.00

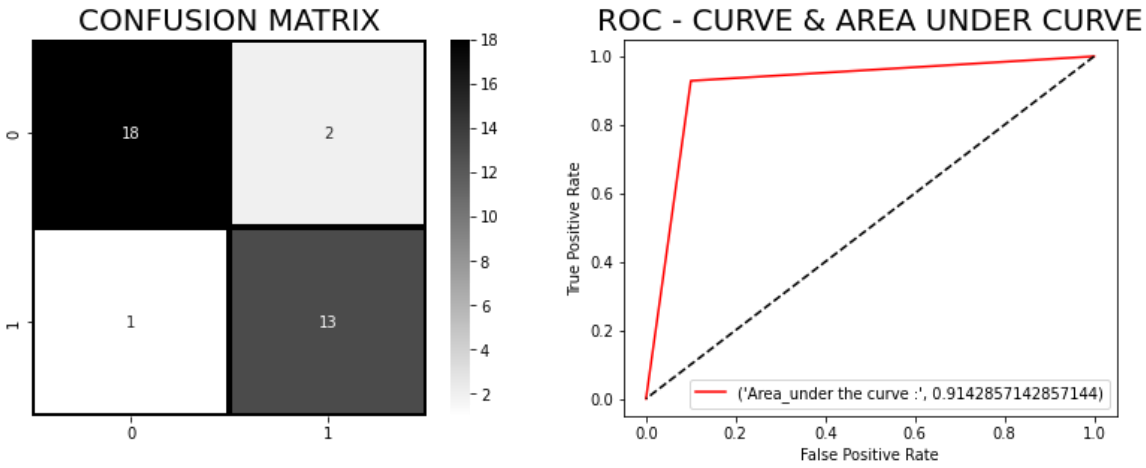

Figure 4.20. Confusion Metric and ROC curve for XG-Boost model (non-PCA)

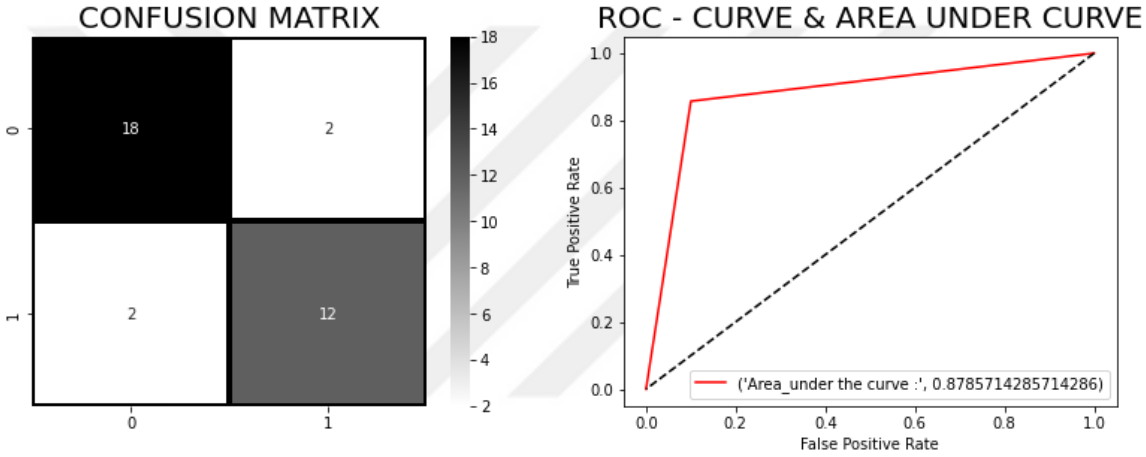

Figure 4.21. Confusion Metric and ROC curve for XG-Boost model (PCA)

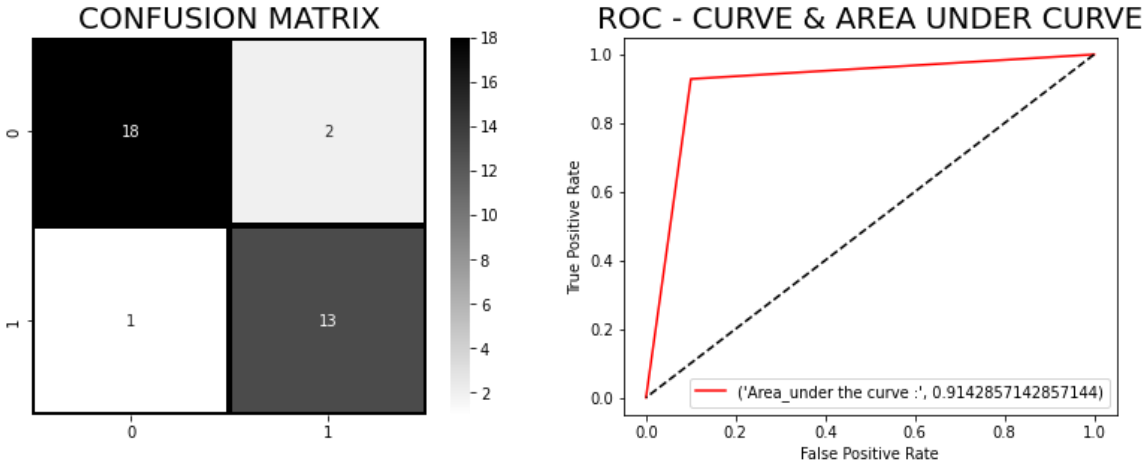

Figure 4.22. Confusion metric and ROC curve of XG-Boost model (RF importance)

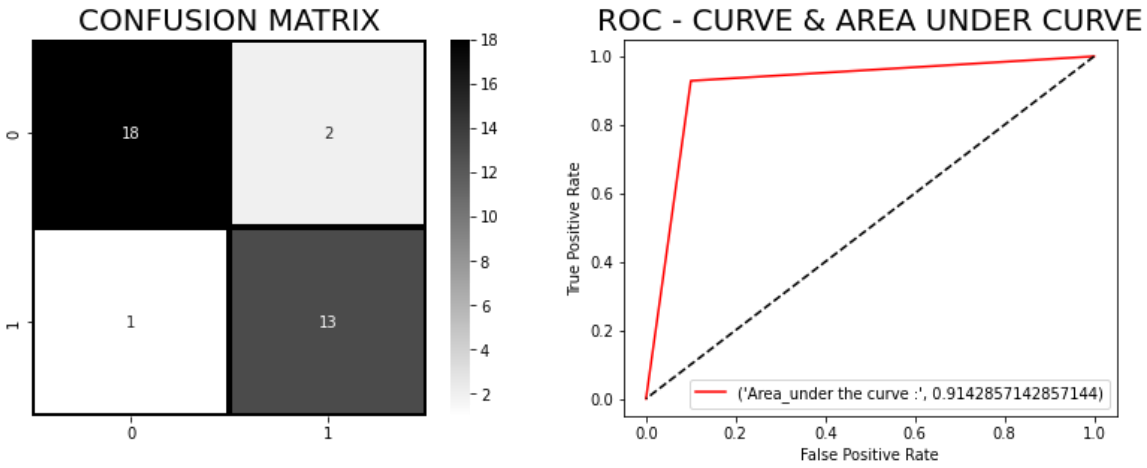

Figure 4.23. Confusion metric and ROC curve of XG-Boost model (lasso regularization)

4.3.7. Results of Gaussian-NB Classifier

Table 4.7 shows classification results obtained from the Gaussian-NB classifier of both non-PCA and PCA versions. The validation accuracy of both PCA and non-PCA versions also obtained 91.17% with ROC curve value of 91.42%. Compared to the three previous models, Gaussian-NB, RF, and DT obtained identical results whether PCA employed or not in the original dataset. But Gaussian-NB performed identically in both versions. Moreover, Figure 4.24 shows the confusion metric of and ROC value of the non-PCA XG-Boost model along with the ROC curve value and confusion matrix of the PCA version illustrated in Figure 4.25.

Moreover, with RF importance, the Gaussian-NB classifier model achieved an accuracy of 76% with a ROC value of 77.85, which is lower than the non-PCA and PCA versions. Figure 4.26 shows the confusion metric of the Gaussian-NB model and ROC curve values with RF importance. Besides, the Lasso Regularization-based Gaussian-NB model is overfitted. In this configuration, models with non-PCA versions (raw dataset) and PCA performed better and achieved similar accuracy. Still, the RF importance-based dataset achieved lower accuracy alongside lasso regularization utilized in the Gaussian-NB model overfitted. Figure 4.27 shows the confusion metric of Gaussian-NB and ROC curve values for the lasso regularization-based Gaussian-NB model.

Table 4.7. Classification report of Gaussian-NB model

Methods	Accuracy %	ROC Value (Avg.)%	Precision (Avg.)%	Recall (Avg.)%	F1-score (Avg.)%
Gaussian-NB (non-PCA)	91.17	91.42	91.00	91.00	91.00
Gaussian-NB (PCA)	91.17	91.42	91.00	91.50	91.00
Gaussian-NB (RF Importance)	76.00	77.85	77.00	78.00	76.00
Gaussian-NB (Lasso Regularization)	100	100	100	100	100

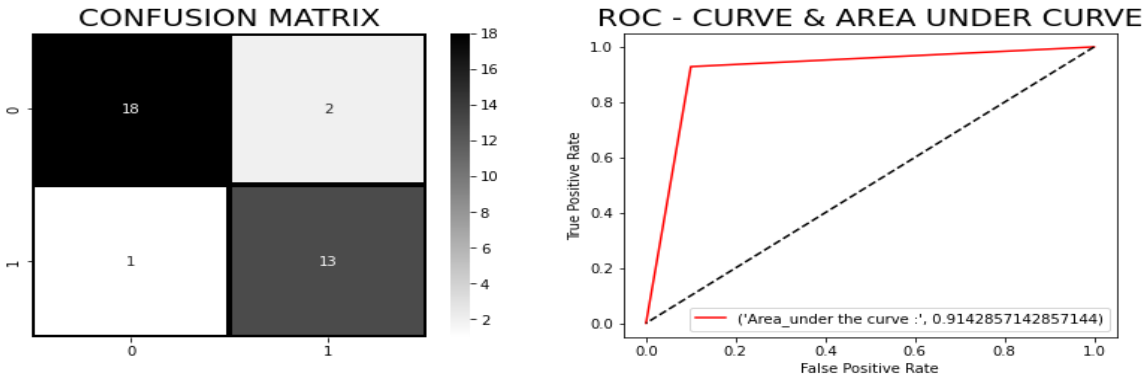

Figure 4.24. Confusion Metric and ROC curve for Gaussian-NB model (non-PCA)

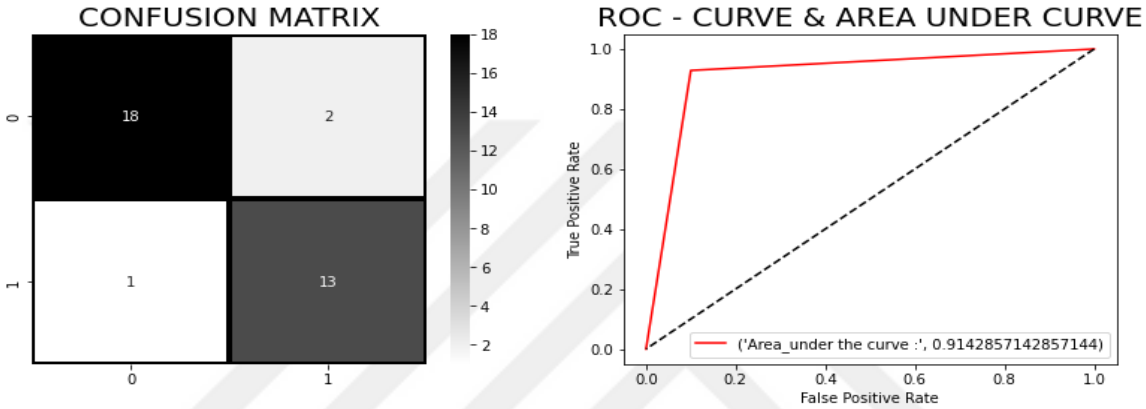

Figure 4.25. Confusion Metric and ROC curve for Gaussian-NB model (PCA)

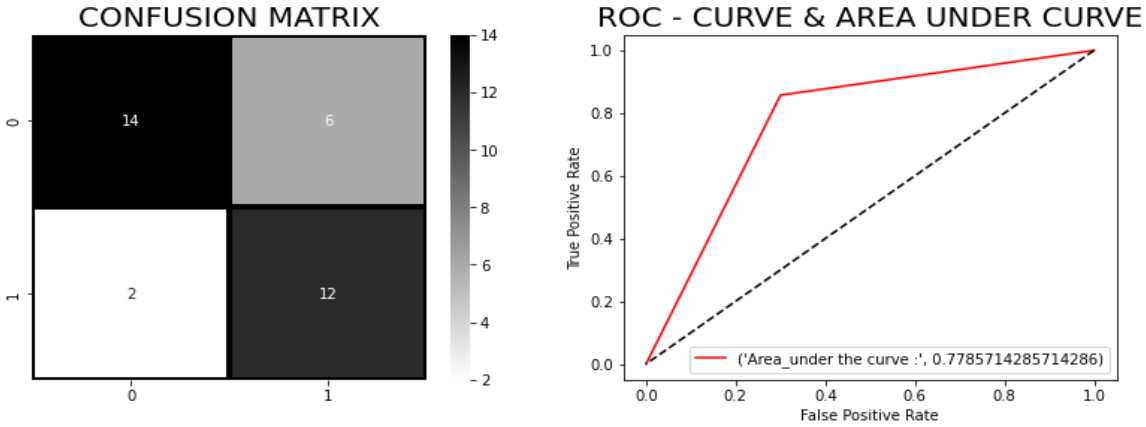

Figure 4.26. Confusion metric and ROC curve of Gaussian-NB model (RF importance)

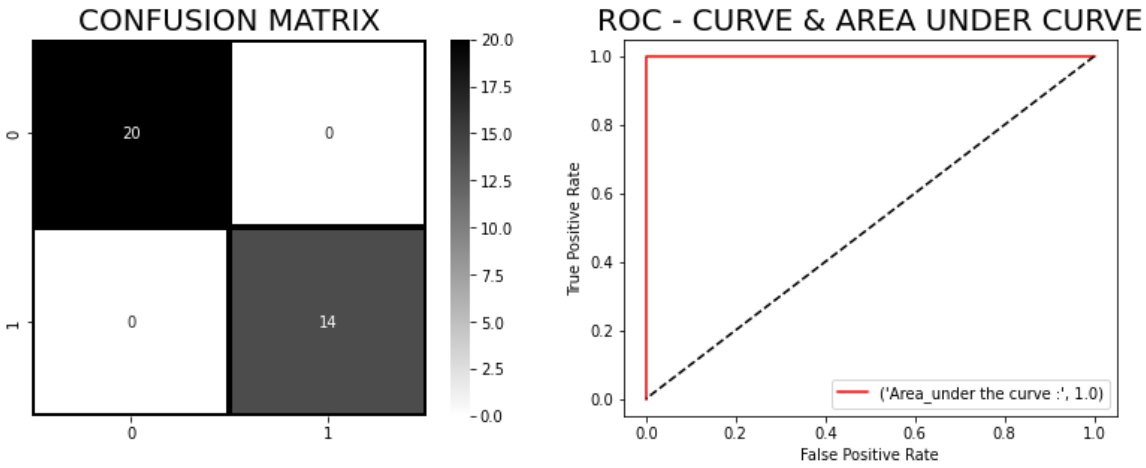

Figure 4.27. Confusion metric and ROC curve of Gaussian-NB model (lasso regularization)

4.4. Results of DL-based Classifier

4.4.1. Results of DNN Classifier

The NN model is utilized and trained with 50epocs with early stopping at epoch 12th because it already obtained 100% accuracy, though the training loss was still high, indicating overfitting. Table 4.8 tabulated the classification results obtained from the DNN model. The DNN model with non-PCA (raw dataset) achieved an accuracy of 58.50% with a ROC curve value of 90, and with PCA, obtained an accuracy of 88.20%, which is better than the non-PCA version. Moreover, Figures 4.28-4.29 shows the confusion metric of the DNN model along with ROC curve value with and non-PCA and PCA versions. Besides, the initial version of the DNN does not seem reasonable as it stated overfitting on the dataset. At the same time, it misclassified many labels.

Similarly, the DNN model, with RF importance, achieved an accuracy of 85% with a ROC value of 84.28. Figure 4.30 shows the confusion metric of the DNN model and ROC curve values for the RF importance-based model. Besides, Lasso Regularization-based DNN model achieved an accuracy of 97%, which is higher than the PCA, non-PCA, and RF importance versions utilized for the DNN model. In this configuration, the DNN model with lasso regularization performed better than the rest. Figure 4.31 shows the confusion metric of DNN and ROC curve values for lasso regularization-based DNN.

Table 4.8. Classification report of NN model

Methods	Accuracy %	ROC Value (Avg.)%	Precision (Avg.)%	Recall (Avg.)%	F1-score (Avg.)%
DNN (non- PCA)	58.50	90.00	85.70	29.00	50.00
DNN (PCA)	88.20	87.85	87.85	88.00	88.00
DNN (RF Importance)	85.00	84.28	85.00	84.00	85.00
DNN (Lasso Regularization)	97.00	96.42	98.00	96.00	97.00

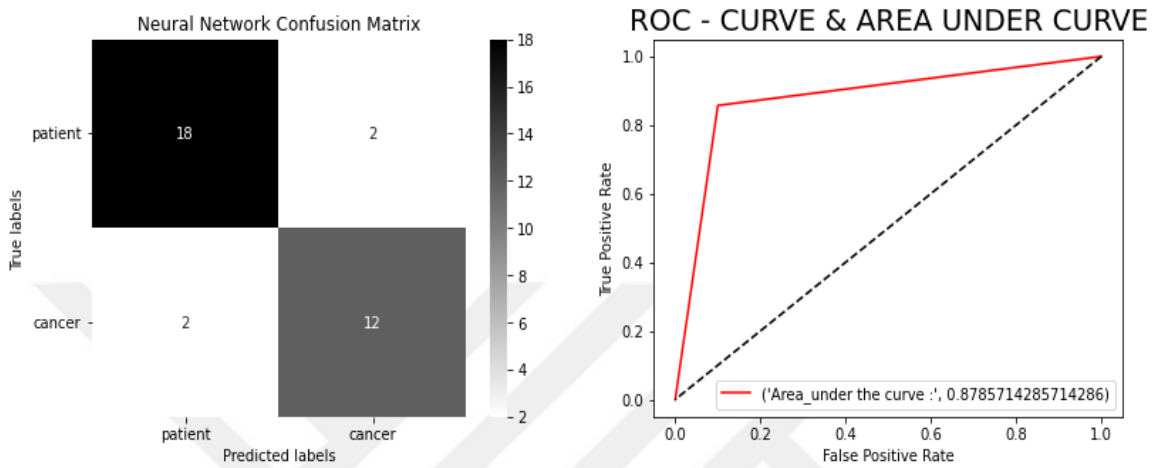

Figure 4.28 Confusion Metric and ROC curve for DNN model (PCA)

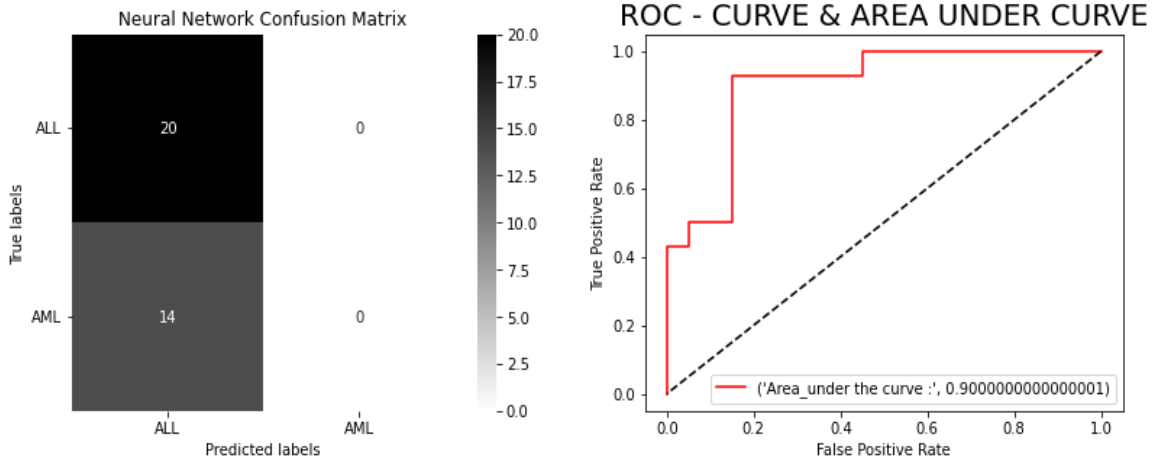

Figure 4.29. Confusion Metric and ROC curve for DNN model (non-PCA)

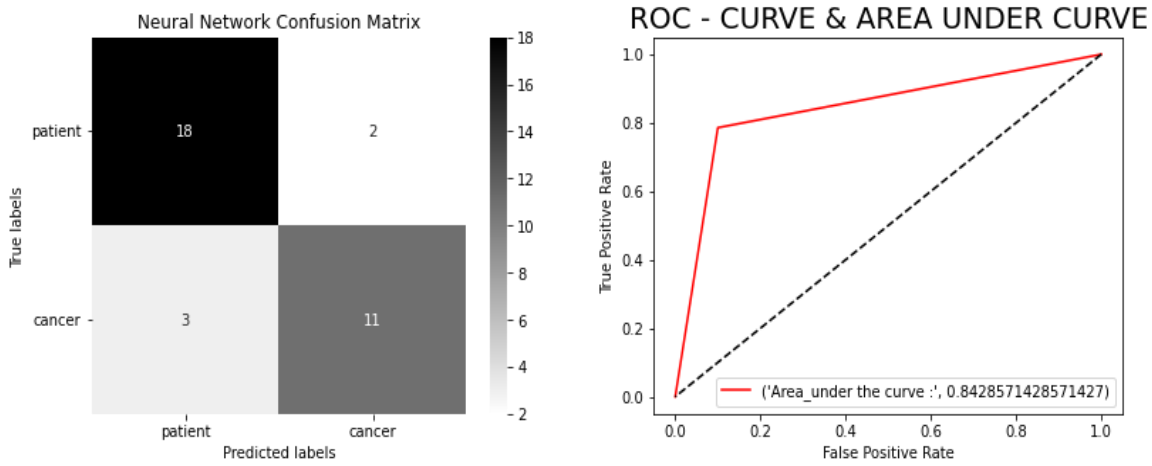

Figure 4.30. Confusion metric and ROC curve of DNN model (RF importance)

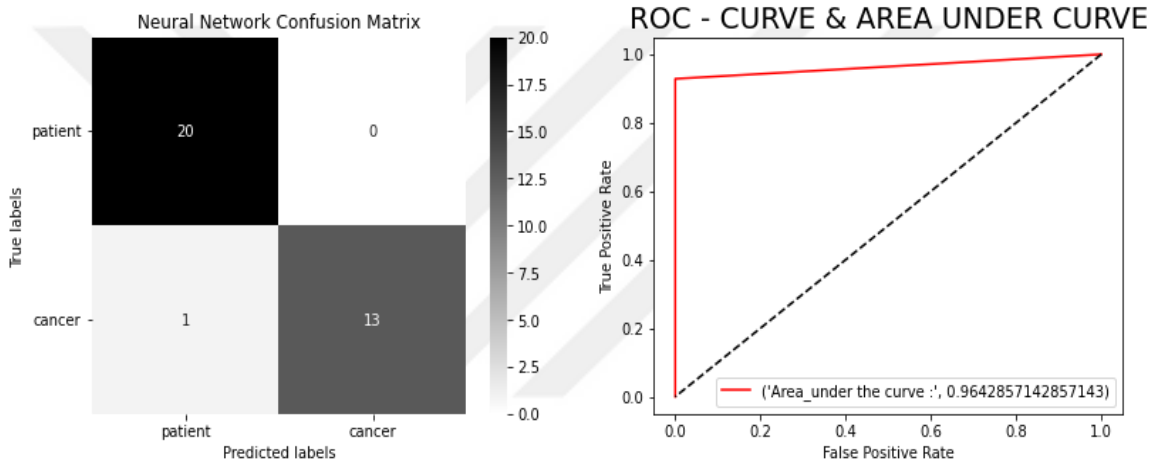

Figure 4.31. Confusion metric and ROC curve of DNN model (lasso regularization)

4.5. Results of Deep-RL Classifier

Deep-RL was also used with six different models with and without utilizing the PCA pre-processing technique. The Learning rate throughout training is set to 0.00025. Moreover, in this formation, both the PCA and non-PCA versions of all models did not perform well; the achieved performance is not on par with other models, but training results were quite good also, and most of the models were overfitted. Figure 4.32 shows the ROC curve values of models 1-6 non-PCA. Lastly, Figure 4.33 shows all six models' confusion metrics. The outcome of the non-PCA version of all six Deep-RL models was not impressive compared to traditional ML classifiers such as LR and SVM performed better.

Furthermore, the non-PCA version of all six Deep-RL model results stated that out of six non-PCA versions of Deep-RL models, model-6 performed better and achieved a test-ROC value of 61. Likewise, the results of the PCA version of six Deep-RL models stated that out of six PCA version of Deep-RL models, models 3, 5, and 6 achieved a test-ROC value of 87.

Figure 4.34 shows the ROC curve values of six models with PCA, and Figure 4.35 illustrates the confusion metric of all six models with PCA.

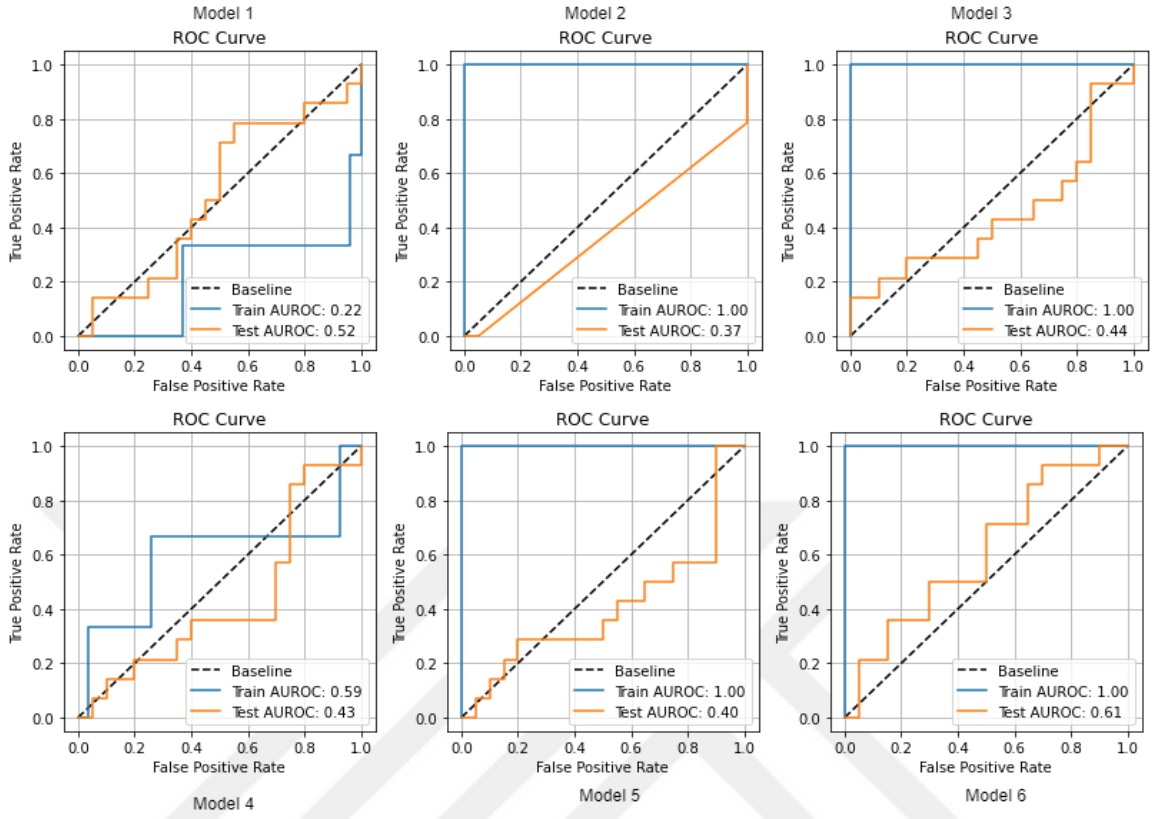

Figure 4.32. ROC curve for Deep-RL models 1-6 (non-PCA)

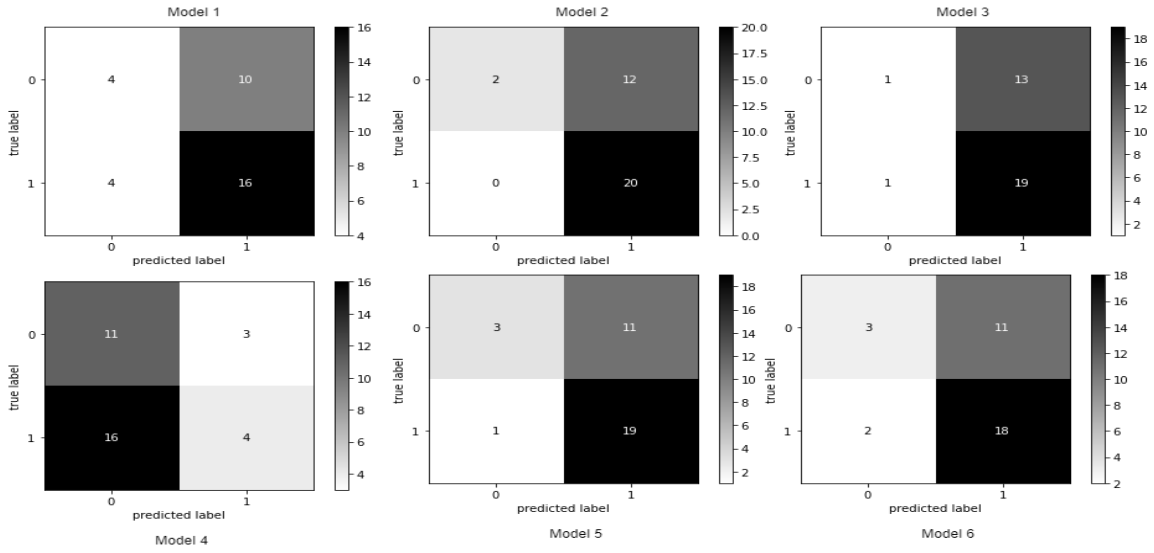

Figure 4.33. Confusion metric of Deep-RL models 1-6 (non-PCA)

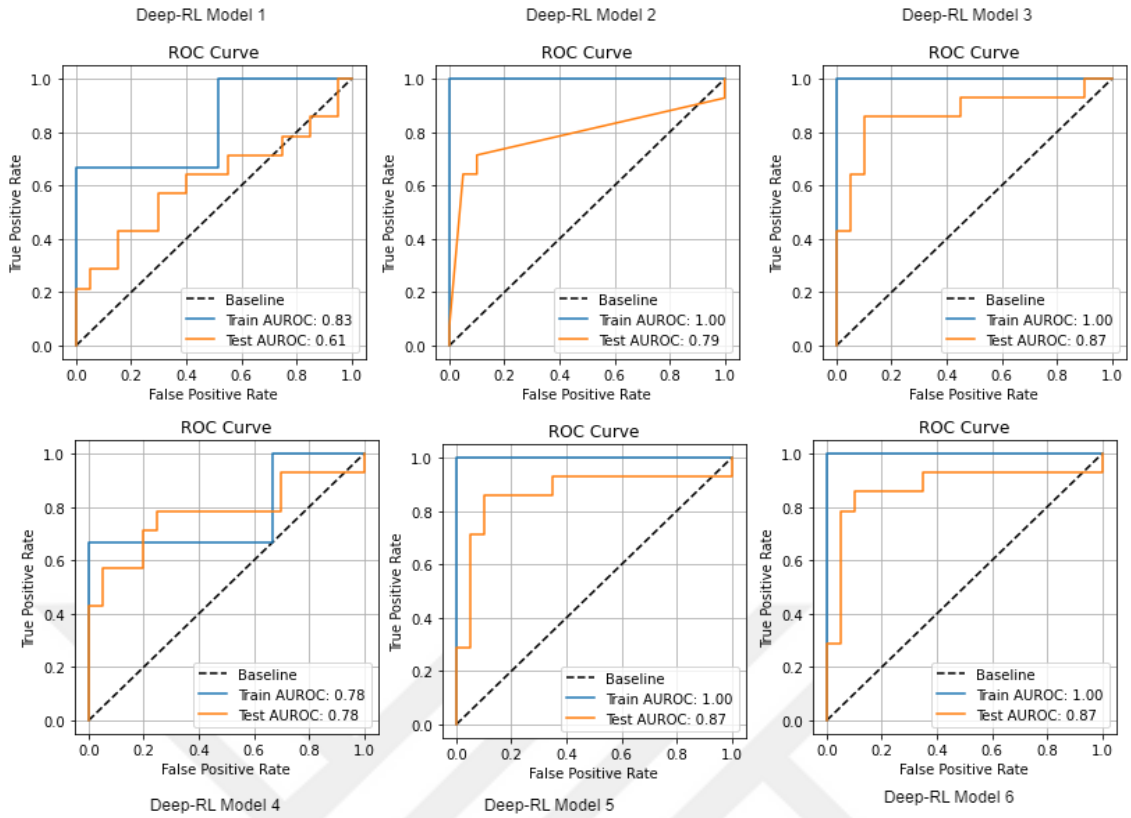

Figure 4.34. ROC curve for Deep-RL models 1-6 (PCA)

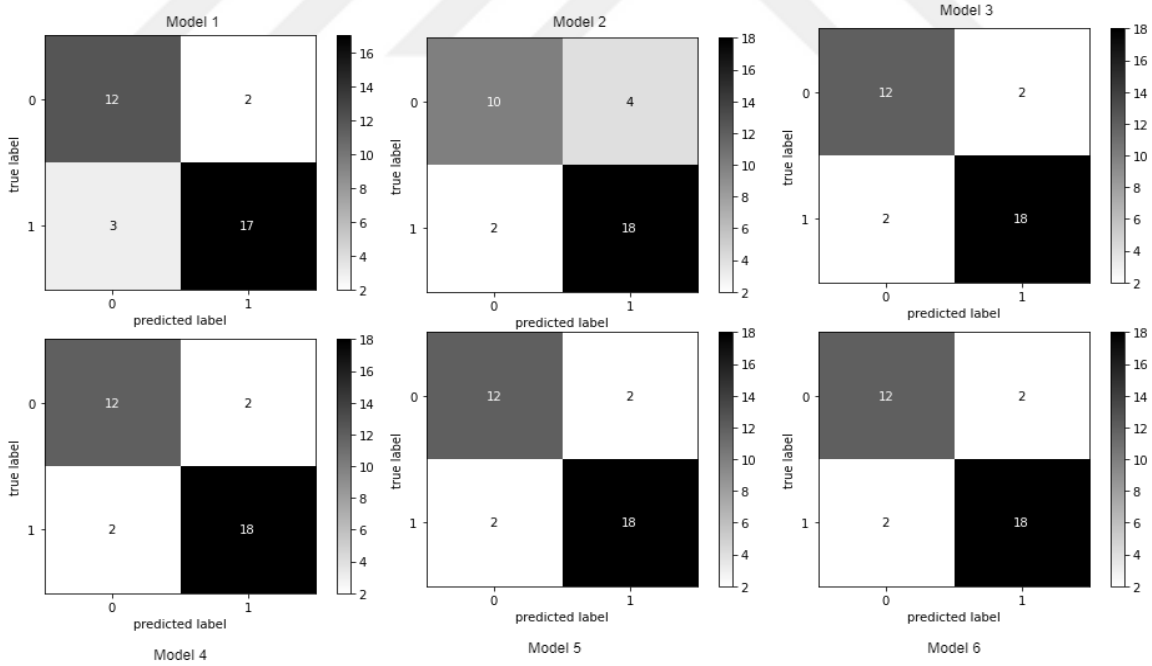

Figure 4.35. Confusion metric of Deep-RL models 1-6 (PCA)

Besides, for the RF importance, picked the model that achieved higher accuracy out of 6 different configured Deep-RL models. The Deep-RL model 6 achieved an accuracy of 84.24%, which is higher than the rest of the five other configurations. Figure 4.36 shows the

confusion metric of the Deep-RL model and ROC curve values with RF importance. Figure 4.37 shows the confusion metric of the Deep-RL model and ROC curve values with lasso regularization.

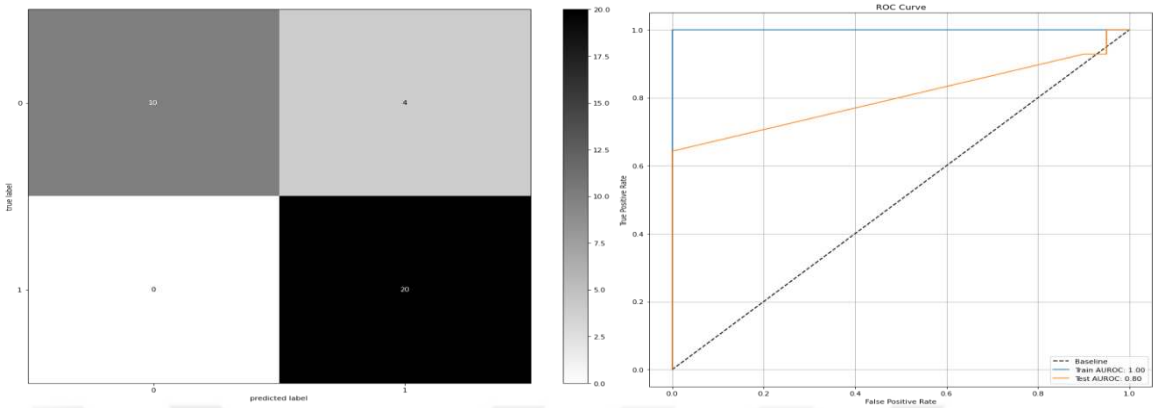

Figure 4.36. Confusion metric and ROC curve for Deep-RL model-6 (RF importance)

However, the Deep-RL with lasso regularization model 1 achieved a higher accuracy of 73.53% but lower accuracy as compared to the best model of Deep-RL with RF importance.

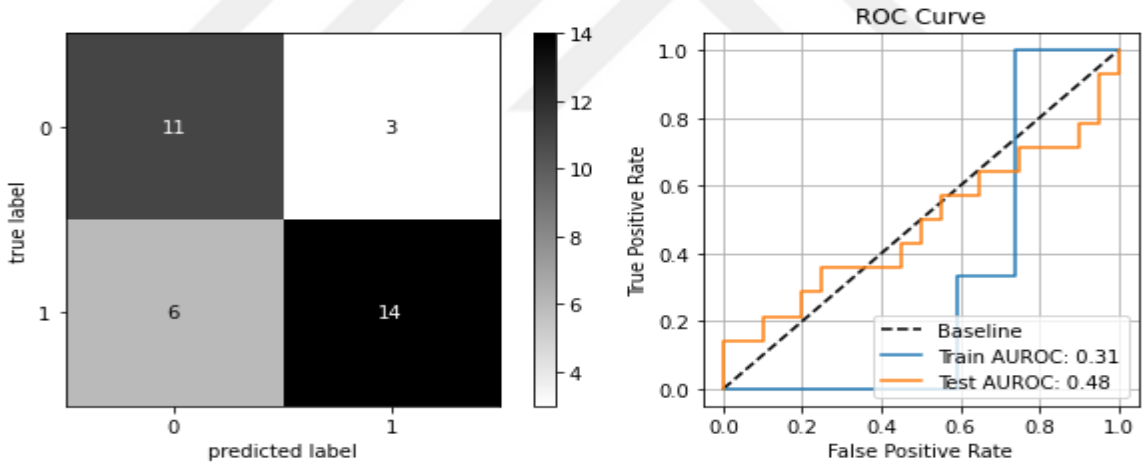

Figure 4.37. Confusion metric and ROC curve for Deep-RL model-1 (lasso regularization)

4.6. Comparison of Classifiers

This thesis evaluated different models with three different configurations of the dataset, and the raw dataset has been examined. Table 4.9 tabulated the accuracies of ML-based models along with the ROC curve values after pre-processing technique (PCA) was applied to models to improve the process of classifying cancer genes based upon microarray gene expression, achieved by utilized models, with the scaled data along with grid search to help fine-tune the model. The SVM models obtained higher accuracy of 91.20%, which is superior to other classifiers. The rest of the model's outcome is illustrated in Figure 4.38.

Table 4.9. Accuracies and ROC values with best and worst percentage of the classifiers (PCA)

Model	Accuracy %	ROC Value%
SVM	91.20	94.30
LR	91.17	91.40
Gaussian-NB	91.17	91.42
DT	88.23	88.90
XG-Boost	88.24	87.90
RF	88.23	87.90
NN	88.20	87.90
K-NN	85.29	85.40

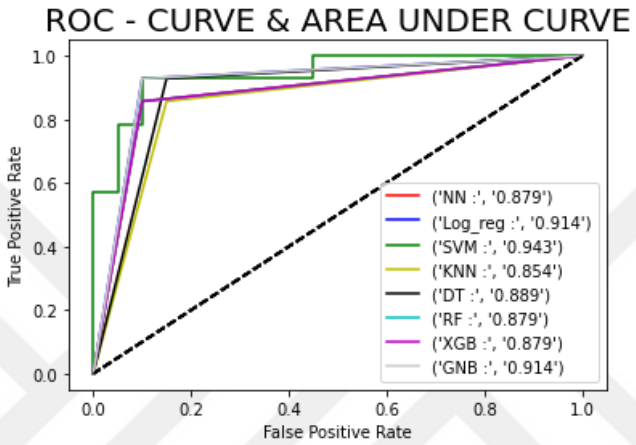

Figure 4.38. ROC curve values of 8 models (PCA)

The non-PCA version of LR achieved higher accuracy of 97% and a ROC value of 97.5%. However, the second-best model, SVM, obtained 94.12% accuracy, but the ROC curve value is better than the LR model. Table 4.10 shows the results of the classifiers. In this non-PCA formation LR model obtained better accuracy as compared to other ML-based models. The ROC curve values of all models except Deep-RL are illustrated in Figure 4.39.

Table 4.10. Accuracies and ROC values with best and worst percentage of the classifiers (non-PCA)

Model	Model Accuracy %	ROC Value%
LR	97.00	97.05
SVM	94.12	99.46
DT	91.18	91.42
RF	91.17	91.42
XG-Boost	91.17	91.42
Gaussian-NB	91.17	91.42
K-NN	82.35	80.71
DNN	58.08	85.07

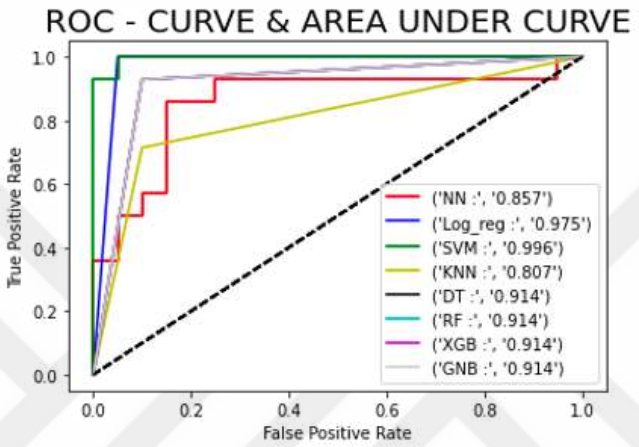

Figure 4.39. ROC curve of 8 models (non-PCA)

Furthermore, comparing the non-PCA versions of the ML-based classifier, the outcomes provide no change in accuracy in any of the models, which means PCA did not affect the outcomes of the models. Besides, the only change in the ROC value of the DNN model was when the non-PCA version of the DNN model obtained a ROC value of 90 as compared to the PCA version of the NN model obtained a ROC value of 88.

Moreover, comparing the results of six different models of Deep-RL that are key purpose of this thesis. Thus, in this formation, both PCA and non-PCA version of six models are examined to select the best. Table 4.12 shows the evaluation of the PCA version for models 1-6 ranked-wise. Figure 4.40 illustrates the classification report of the PCA version of Deep-RL models 1-6. The models are ranked to show the best model out of six. Meanwhile, three models (3,5,6) performed identically in every aspect and ranked best amongst other models.

Table 4.11. Evaluation of Deep-RL models with PCA (1-6)

Models (PCA)	Accuracy%	Precision%	Recall%	F1 Score%	ROC%
Deep-RL Model 3	88.24	85.71	85.71	85.71	87.00
Deep-RL Model 5	88.24	85.71	85.71	85.71	87.00
Deep-RL Model 6	88.24	85.71	85.71	85.71	87.00
Deep-RL Model 4	35.29	16.67	14.29	15.39	78.00
Deep-RL Model 2	82.35	83.33	71.43	76.92	79.00
Deep-RL Model 1	85.30	80.00	85.71	82.76	61.00

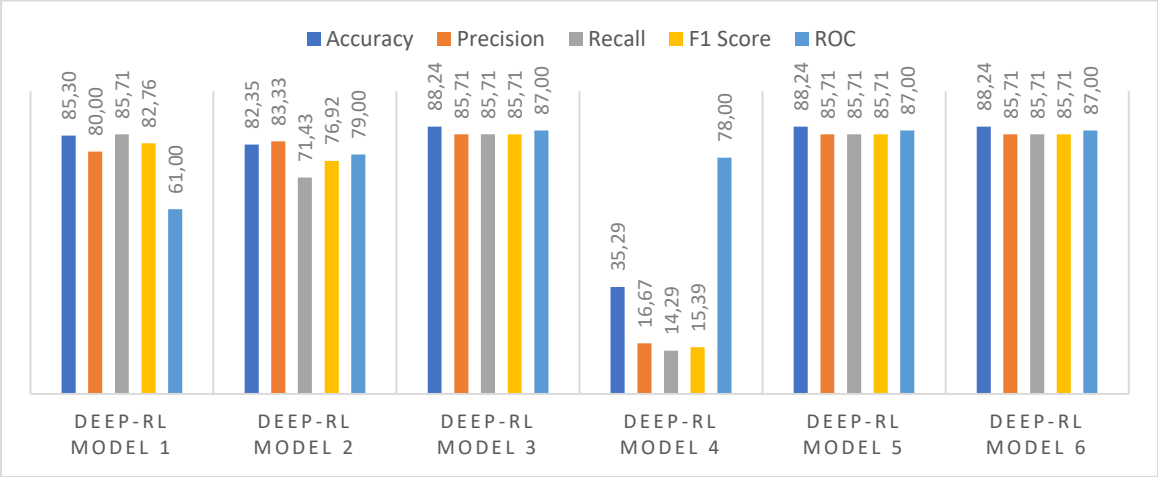

Figure 4.40. Evaluation metrics of Deep-RL 1-6 models (PCA)

Besides, the evaluation metrics of all six non-PCA versions of Deep-RL models are illustrated in Figure 4.41. Meanwhile, three model-6 performed better, followed by model-1, considered best among other models. Table 4.12 shows the classification report of the non-PCA version for models 1-6, respectively. The models are ranked to show the best of the six models in the non-PCA version.

Table 4.12. Evaluation of Deep-RL models non-PCA (1-6)

Models (PCA)	Accuracy%	Precision%	Recall%	F1 Score%	ROC%
Deep-RL Model 3	88.24	50.00	71.42	12.05	44.00
Deep-RL Model 5	88.24	75.00	21.42	33.33	40.00
Deep-RL Model 6	88.24	60.00	21.42	31.57	61.00
Deep-RL Model 4	35.29	40.74	78.57	53.65	43.00
Deep-RL Model 2	82.35	100.00	14.28	25.00	37.00
Deep-RL Model 1	85.30	50.00	28.57	36.36	52.00

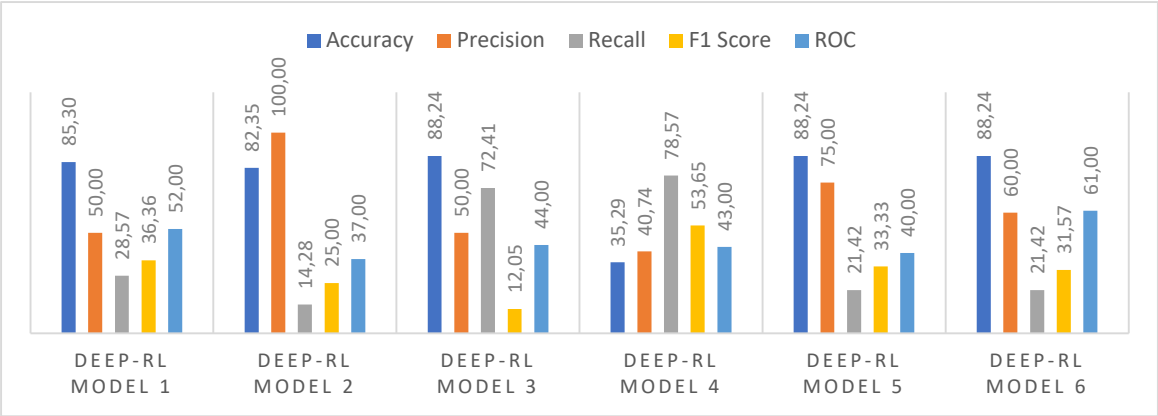

Figure 4.41. Evaluation metrics of Deep-RL 1-6 models (non-PCA)

Furthermore, with ML and DL-based models with the RF importance configuration, compared the accuracies of models along with ROC values are tabulated in Table 4.13. in this configuration, SVM models obtained higher accuracy of 97%, which is superior to other classifiers. ROC values of all models are illustrated in Figure 4.42.

Table 4.13. Accuracies and ROC values of the classifiers (RF importance)

Model	Accuracy %	Precision %	Recall %	F1 Score %	ROC %
LR	85.00	85.00	86.00	85.00	86.42
SVM	**97.00**	**97.00**	**97.00**	**97.00**	**97.50**
K-NN	94.00	94.00	94.00	94.00	93.92
DT	91.00	91.00	91.00	91.00	91.42
RF	91.00	91.00	91.00	91.00	91.42
XGBoost	91.00	91.00	91.00	91.00	91.42
Gaussian NB	76.00	77.00	78.00	76.00	77.85
DNN	85.00	85.00	84.00	85.00	84.28

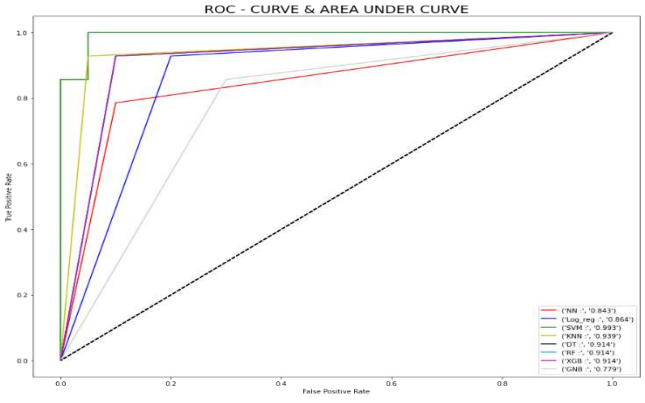

Figure 4.42. ROC curve of 8 models (RF importance)

Moreover, ML and DL models with lasso regularization compared the accuracies of models with ROC values, tabulated in Table 4.14. Three models, LR, Gaussian-NB, and SVM models overfitted. However, DNN achieved higher accuracy than the rest of the models. The ROC curve values of all models except Deep-RL are illustrated in Figure 4.43.

Table 4.14. Accuracies and ROC values of the classifiers (lasso regularization)

Model	Accuracy %	Precision %	Recall %	F1 Score %	ROC %
LR	100.00	100.00	100.00	100.00	100.00
SVM	100.00	100.00	100.00	100.00	100.00
K-NN	91.00	93.00	89.00	91.00	89.28
DT	91.00	91.00	91.00	91.00	91.42
RF	91.00	91.00	91.00	91.00	91.42
XGBoost	91.00	91.00	91.00	91.00	91.42
Gaussian NB	100.00	100.00	100.00	100.00	100.00
DNN	**97.00**	**98.00**	**96.00**	**97.00**	**96.42**

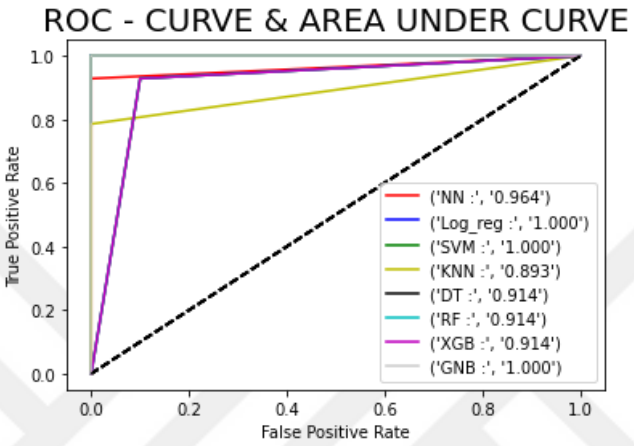

Figure 4.43. ROC curve of 8 models (lasso regularization)

Additionally, for the Deep-RL model dataset based on RF importance and lasso regularization. In this formation, selected the best-performed configuration of 6 different configurations. In that case, the Deep-RL model 6 with the RF importance dataset performed better than the rest of the configurations, which obtained an accuracy of 84.24%. Besides the Deep-RL model with the lasso regularization dataset, model 1 achieved the highest accuracy of 73.53%. Table 4.15 shows the results of both models with the highest achieved accuracies along with other data. Also, Figure 4.44 shows the confusion metric and ROC curve value for Deep-RL model 6 with the RF importance dataset, and Figure 4.45 shows the confusion metric and ROC curve value for Deep-RL model 1 with the lasso regularization dataset.

Table 4.15. Deep-RL models with higher accuracy of both RF importance and lasso regularization

Model	Accuracy %	Precision%	Recall %	F1 Score %	ROC %
Deep-RL Model 6 (RF importance)	88.24	100.00	71.42	83.30	80.00
Deep-RL Model 1 (lasso regularization)	73.53	64.70	78.57	70.96	48.00

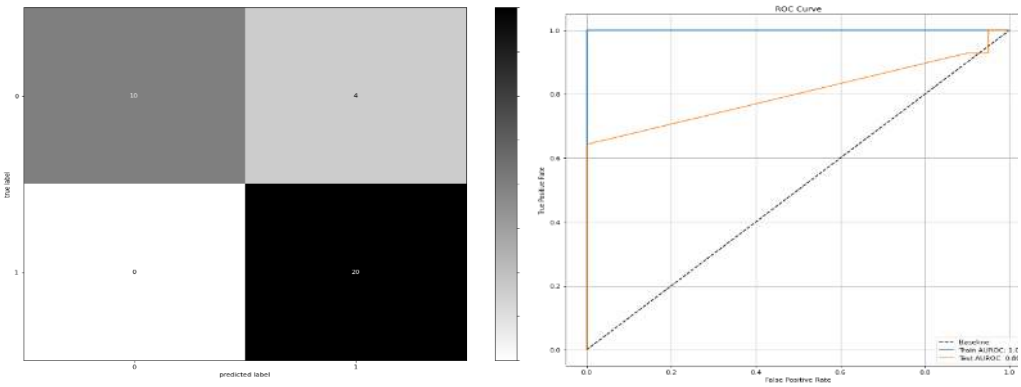

Figure 4.44. Confusion metric and ROC curve for Deep-RL model 6 with RF importance

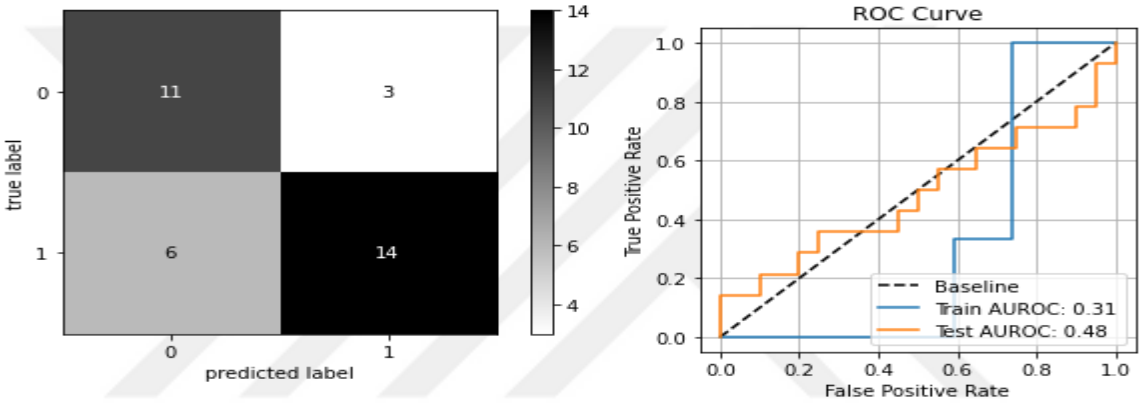

Figure 4.45. Confusion metric and ROC curve for Deep-RL model 1 with lasso regularization

5. DISCUSSION

This study utilized different ML algorithms, such as LR, SVM, DT, RF, Gaussian-NB, and XG-Boost, alongside DL-based algorithms, namely NN, and Deep-RL to identify leukemia cancer subclasses using DNA Gene Expression dataset. Four configurations were utilized for each model of the ML-based algorithm to carry out the experiments. The first one is the PCA version, where PCA is used for dimensionality reduction. Second is every model's non-PCA (raw dataset) version to compare whether PCA affects the outcomes. The third is a dataset with a RF feature selection algorithm utilized with every ML-based classifier, DNN classifier, and 6 different configurations of the Deep-RL model. Fourth is the dataset with lasso regularization—similarly, each ML classifier, DNN classifier, and Deep-RL model with six different configurations. We compared six other formations for the Deep-RL with RF feature selection and lasso regularization and selected one configuration that performed better than the rest.

LR and SVM models performed better than ML-based models. SVM obtained an accuracy of 91.20 % with PCA and 94.12% with the raw dataset, respectively, and LR obtained 97.17% with PCA and 97% with the non-PCA version. Table 5.1 shows the accuracy comparison of all four dataset configurations for both ML and DL models except the Deep-RL models.

Table 5.1. Accuracy comparison of ML and DL models

Models	**Non-PCA**	**PCA**	**RF importance**	**Lasso regularization**
	Accuracy%	**Accuracy%**	**Accuracy%**	**Accuracy%**
SVM	94.12	91.20	97.00	100
LR	97.00	91.17	85.00	100
Gaussian-NB	91.17	91.17	76.00	100
DT	91.18	88.23	91.00	91.00
XG-Boost	91.17	88.24	91.00	91.00
RF	91.17	88.23	91.00	91.00
DNN	58.50	88.20	85.00	97.00
K-NN	82.35	85.29	94.00	91.00

Additionally, three dataset versions of three ML-based models, DT, XGBoost, and RF, obtained identical accuracy and ROC curve values. However, the PCA version of four ML-based models, DT, XG-Boost, RF, and Gaussian-NB, obtained identical results. The lasso regularization version of the DNN model obtained an accuracy of 97%, the highest accuracy obtained in that formation, and the PCA version obtained 88.20%, which is lower than the SVM, LR, and Gaussian-NB but on par with another ML-based model. The non-PCA version obtained the worst accuracy among all the other utilized models.

Similarly, for Deep-RL, six different models were used to compare the outcomes. Moreover, all six models of Deep-RL are also utilized with both PCA and non-PCA versions. In this formation, the PCA version of three models (3,5,6) obtained higher accuracy. However, in the formation of non-PCA, add updated accuracies like PCA. Table 5.2 shows the accuracy comparison of both PCA and non-PCA versions of Deep-RL models. Moreover, comparing the results of both PCA and non-PCA versions, we can see that the PCA version got much better and improved outcomes than the non-PCA version.

Table 5.2. Best models with PCA and non-PCA dataset

Models	**PCA**	**Non-PCA**
	Accuracy%	**Accuracy%**
Deep-RL Model 3	88.24	58.82
Deep-RL Model 5	88.24	64.7
Deep-RL Model 6	88.24	61.76

Furthermore, of the six Deep-RL models, model 6 utilizing the RF importance achieved the highest accuracy. In the lasso regularization dataset version of 6 different Deep-RL models, model 1 achieved a higher accuracy than the rest. The accuracy comparison of both formations is tabulated in Table 5.3.

Table 5.3. The best model with RF importance and lasso regularization dataset

Deep-RL	**Accuracy%**	**Features Selection Models**
Model 6	88.24	Randon Forest Importance
Model 1	73.53	Lasso Regularization

6. CONCLUSION

Numerous researchers worked on analyzing biological data and solutions related to these problems. Different machine-learning methods have become more and more popular in this area. Multiple DNA microarray data in practice are characterized by very high dimensional vectors, which takes great obstacles in data mining and further processing. Besides increasing the learning cost, high dimensionality also declines the learning performance, entitled the "curse of dimensionality". Consequently, dimensionality reduction has greatly benefited numerous fields and their applications, such as microarray data analysis.

Under this framework, this thesis compares the performances of 8 different models and six Deep-RL models by utilizing the gene expression dataset. In this proposed method, the initial set of categorization options were 8 other ML-based models for leukemia cancer classification based on gene expression.

This framework evaluated different models with PCA utilization to reduce the dimensionality of the dataset along with non-PCA versions of the model that have been examined. So far, LR and SVM models have proved to be effective methods for classifying AML and ALL from the rest of the proposed ML-based methods, along with Deep-RL, which also did not perform well. LR and SVM models obtained an accuracy of 97% in both PCA and non-PCA formation, which is superior to other classifiers.

However, three models, K-NN, DT, and RF have identical model accuracy and ROC curve values. Like those three mentioned models, the Gaussian-NB model also obtained comparable results with no change in whether PCA was employed in the original dataset. The validation accuracy of both PCA and non-PCA versions of all these four models, namely K-NN, DT, and RF, Gaussian-NB, obtained 91.17% with a ROC curve of 91.42%. In contrast, the NN model achieved an accuracy of 58.8% with a ROC curve value of 88.21 with PCA and without PCA version only change in ROC curve value of 89.28 with the same model accuracy. Surprisingly, NN did not perform quite well in classification tasks because it usually requires much more data than traditional ML algorithms. That's why Gaussian-NB performed better because naive Bayes deals much better with little data.

In forming the RF importance dataset for ML and DL models, The SVM performed better and achieved an accuracy of 97% with a ROC value of 97.50. The k-NN model obtained the second-highest accuracy of 94% with a ROC value of 93.92. On the other hand, Gaussian-NB received the lowest accuracy of 76%. In the case of Deep-RL 6 different configurations, Model 6 obtained a higher accuracy of 84.24%.

In form, the lasso regularization dataset for ML and DL models, three ML models, Gaussian-NB, LR, and SVM, overfitted. The DL-based model DNN achieved a higher accuracy

of 97%. Alternatively, the lasso regularization dataset for six different Deep-RL configurations, model 1, obtained a higher accuracy of 73.53%.

We can see that if we forecast that every patient has AML, we will often be accurate, but all the models provided ineffective results for ALL predictions. A grid search using KNN is also overfit. The only difference is the results of a randomized search are better to some extent. The difficulty is that grid and random search employ the model that best matches the training data without considering how well it will generalize to the test data. Overall, the model is inaccurate, does not generalize to the test data, and has a significant variance issue. The Deep-RL did not add anything regarding results, as all six tested models provided highly overfitted results. The overall experiment showed that the data is inefficient for accurate prediction; GAN-based or real data is needed to improve the results.

REFERENCES

Ablin, A. R. (1972). The treatment of acute myelocytic leukemia. *California Medicine*, *117*(6), 62–63.

Alamgir Sarder, Md., Maniruzzaman, Md., & Ahammed, B. (2020). Feature Selection and Classification of Leukemia Cancer Using Machine Learning Techniques. *Machine Learning Research*, *5*(2), 18. https://doi.org/10.11648/j.mlr.20200502.11

Ali, I., Hart, G. R., Gunabushanam, G., Liang, Y., Muhammad, W., Nartowt, B., Kane, M., Ma, X., & Deng, J. (2018). Lung Nodule Detection via Deep Reinforcement Learning. *Frontiers in Oncology*, *8*, 108. https://doi.org/10.3389/fonc.2018.00108

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient Machine Learning for Big Data: A Review. *Big Data Research*, *2*(3), 87–93. https://doi.org/10.1016/j.bdr.2015.04.001

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, 6.

Awad, M., & Khanna, R. (2015). *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Apress. https://doi.org/10.1007/978-1-4302-5990-9

Azar, A. T. (2013). Fast neural network learning algorithms for medical applications. *Neural Computing and Applications*, *23*(3–4), 1019–1034. https://doi.org/10.1007/s00521-012-1026-y

Balaprakash, P., Egele, R., Salim, M., Wild, S., Vishwanath, V., Xia, F., Brettin, T., & Stevens, R. (2019). Scalable Reinforcement-Learning-Based Neural Architecture Search for Cancer Deep Learning Research. *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–33. https://doi.org/10.1145/3295500.3356202

Bolshakova, N., Azuaje, F., & Cunningham, P. d. (2005). An integrated tool for microarray data clustering and cluster validity assessment. *Bioinformatics*, *21*(4), 451–455. https://doi.org/10.1093/bioinformatics/bti190

Bowyer, K., Kopans, D., Kegelmeyer, W.P., Moore, R., Sallam, M., Chang, K. and Woods, K. (1996). The digital database for screening mammography. In *In Third international workshop on digital mammography* (Vol. 58).

Boyle, P., Levin, B., International Agency for Research on Cancer, & World Health Organization (Eds.). (2008). *World cancer report 2008*. International Agency for Research on Cancer ; Distributed by WHO Press.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., & others. (2015). Xgboost: Extreme gradient boosting. *R Package Version 0.4-2*, *1*(4), 1–4.

Chih, C. C., & Chih, J. L. (2022). *LIBSVM -- A Library for Support Vector Machines*. https://www.csie.ntu.edu.tw/~cjlin/libsvm/

Childhood Leukemia Early Detection, Diagnosis, and Types. (2022). https://www.cancer.org/cancer/leukemia-in-children/detection-diagnosis-staging.html

Cohn, J. F., Jeni, L. A., Onal Ertugrul, I., Malone, D., Okun, M. S., Borton, D., & Goodman, W. K. (2018). Automated Affect Detection in Deep Brain Stimulation for Obsessive-Compulsive Disorder: A Pilot Study. *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 40–44. https://doi.org/10.1145/3242969.3243023

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines: And other kernel-based learning methods*. Cambridge University Press.

Cunningham, P. (2007). *K-Nearest Neighbour Classifiers, University College Dublin, Padraig. Cunningham@ ucd. Ie, Dublin Institute of Technology Sarahjane. Delany@ comp. Dit*. ie, Technical Report UCD-CSI-2007-4.

de Carvalho Couto, C., Freitas-Silva, O., Morais Oliveira, E. M., Sousa, C., & Casal, S. (2022). Near-Infrared Spectroscopy Applied to the Detection of Multiple Adulterants in Roasted and Ground Arabica Coffee. *Foods*, *11*(1), Article 1. https://doi.org/10.3390/foods11010061

De Guia, J. M., Devaraj, M., & Vea, L. A. (2019). Cancer classification of gene expression data using machine learning models. *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management, HNICEM 2018*, 1–6. https://doi.org/10.1109/HNICEM.2018.8666435

Deb, K., & Raji Reddy, A. (2003). Reliable classification of two-class cancer data using evolutionary algorithms. *Biosystems*, *72*(1–2), 111–129. https://doi.org/10.1016/S0303-2647(03)00138-2

Devi Arockia Vanitha, C., Devaraj, D., & Venkatesulu, M. (2014). Gene expression data classification using Support Vector Machine and mutual information-based gene selection. *Procedia Computer Science*, *47*(C), 13–21. https://doi.org/10.1016/j.procs.2015.03.178

Di Cataldo, S., & Ficarra, E. (2017). Mining textural knowledge in biological images: Applications, methods and trends. In *Computational and Structural Biotechnology Journal* (Vol. 15, pp. 56–67). https://doi.org/10.1016/j.csbj.2016.11.002

Ehrenstein, V., Nielsen, H., Pedersen, A. B., Johnsen, S. P., & Pedersen, L. (2017). Clinical epidemiology in the era of big data: New opportunities, familiar challenges. *Clinical Epidemiology*, *Volume 9*, 245–250. https://doi.org/10.2147/CLEP.S129779

Facts Spring. (2014). *American Cancer Society, "facts spring 2014 Lymphoma Society: Fighting Blood Cancer, Revised April 2014.* https://www.lls.org/sites/default/files/file_assets/facts.pdf

Francois-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., & Pineau, J. (2018). An Introduction to Deep Reinforcement Learning. *Foundations and Trends® in Machine Learning*, *11*(3–4), 219–354. https://doi.org/10.1561/2200000071

Frigui, H., & Gader, P. (2009). Detection and Discrimination of Land Mines in Ground-Penetrating Radar Based on Edge Histogram Descriptors and a Possibilistic K-Nearest Neighbor Classifier. *IEEE Transactions on Fuzzy Systems*, *17*(1), 185–199. https://doi.org/10.1109/TFUZZ.2008.2005249

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

Ghorbanzadeh, O., Shahabi, H., Mirchooli, F., Valizadeh Kamran, K., Lim, S., Aryal, J., Jarihani, B., & Blaschke, T. (2020). Gully erosion susceptibility mapping (GESM) using machine learning methods optimized by the multi-collinearity analysis and K-fold cross- validation. *Geomatics, Natural Hazards and Risk*, *11*, 1653–1678. https://doi.org/10.1080/19475705.2020.1810138

Gibney, E. (2016). Go players react to computer defeat. *Nature*, nature.2016.19255. https://doi.org/10.1038/nature.2016.19255

Göhlmann, H., & Talloen, W. (2017). *Gene expression studies using affymetrix microarrays*. CRC Press.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, *286*(5439), 531–537. https://doi.org/10.1126/science.286.5439.531

Gorin, N. C. (1998). Autologous Stem Cell Transplantation in Acute Myelocytic Leukemia. *Blood*, *92*(4), 1073–1090. https://doi.org/10.1182/blood.V92.4.1073

Hamada, M., Tanimu, J. J., Hassan, M., Kakudi, H. A., & Robert, P. (2021). Evaluation of Recursive Feature Elimination and LASSO Regularization-based optimized feature

selection approaches for cervical cancer prediction. *2021 IEEE 14th International Symposium on Embedded Multicore/Many-Core Systems-on-Chip (MCSoC)*, 333–339. https://doi.org/10.1109/MCSoC51149.2021.00056

Han, J., & Kamber, M. (2012). *Data mining: Concepts and techniques* (3rd ed). Elsevier.

Hatem, M., & Abdessemed, F. (2009). *Simulation of the Navigation of a Mobile Robot by the QLearning using Artificial Neuron Networks. 547*. https://www.researchgate.net/publication/221551556_Simulation_of_the_Navigation_of_a_Mobile_Robot_by_the_QLearning_using_Artificial_Neuron_Networks

Herland, M., Khoshgoftaar, T. M., & Wald, R. (2014). A review of data mining using big data in health informatics. *Journal Of Big Data*, *1*(1), 2. https://doi.org/10.1186/2196-1115-1-2

Hooshmand, A. (2020). *Machine Learning Against Cancer: Accurate Diagnosis of Cancer by Machine Learning Classification of the Whole Genome Sequencing Data* (arXiv:2009.05847). arXiv. http://arxiv.org/abs/2009.05847

Huang, P. Y., Best, O. G., Almazi, J. G., Belov, L., Davis, Z. A., Majid, A., Dyer, M. J., Pascovici, D., Mulligan, S. P., & Christopherson, R. I. (2014). Cell surface phenotype profiles distinguish stable and progressive chronic lymphocytic leukemia. *Leukemia & Lymphoma*, *55*(9), 2085–2092. https://doi.org/10.3109/10428194.2013.867486

Huljanah, M., Rustam, Z., Utama, S., & Siswantining, T. (2019). Feature Selection using Random Forest Classifier for Predicting Prostate Cancer. *IOP Conference Series: Materials Science and Engineering*, *546*(5), 052031. https://doi.org/10.1088/1757-899X/546/5/052031

Illumina. (2009). *Gene Expression Analysis*. https://www.illumina.com/content/dam/illumina-marketing/documents/products/guides/2009_product_guide_gene_expression_analysis.pdf

Im, D. J., Kim, C. D., Jiang, H., & Memisevic, R. (2016). *Generating images with recurrent adversarial networks*.

Ishan, Shah. (2020, February 13). *Introduction to XGBoost in Python*. Quantitative Finance & Algo Trading Blog by QuantInsti. https://blog.quantinsti.com/xgboost-python/

Jiang, Y., Deng, Z., Chung, F. L., Wang, G., Qian, P., Choi, K. S., & Wang, S. (2017). Recognition of Epileptic EEG Signals Using a Novel Multiview TSK Fuzzy System. *IEEE Transactions on Fuzzy Systems*, *25*(1), 3–20. https://doi.org/10.1109/TFUZZ.2016.2637405

Joachims, T. (1999). 11 making large-scale support vector machine learning practical. In *Advances in kernel methods: Support vector learning* (p. 169). MIT press.

Karim, M. R., Cochez, M., Beyan, O., Decker, S., & Lange, C. (2019). *OncoNetExplainer: Explainable Predictions of Cancer Types Based on Gene Expression Data* (arXiv:1909.04169). arXiv. http://arxiv.org/abs/1909.04169

Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., & Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, *7*(6), 673–679. https://doi.org/10.1038/89044

Khan, S., Sarker, B., & SDGs, P. B. (2019). Application of Supervised Learning Methods for Gene Expression Microarray Data Classification. *Data Science and SDGs: Challenges, Opportunities and Realities*.

Kocatürk, C., CandemiR, C., & Kocabaş, İ. (2022). A Comparative Study of Automatic Detection of Acute Lymphocytic Leukemia with Machine Learning Methods. *Deu Muhendislik Fakultesi Fen ve Muhendislik*, *24*(72), 1021–1032. https://doi.org/10.21205/deufmd.2022247229

Koo, J., Mendiratta, V. B., Rahman, M. R., & Walid, A. (2019). Deep reinforcement learning for network slicing with heterogeneous resource requirements and time varying traffic dynamics. *2019 15th International Conference on Network and Service Management (CNSM)*, 1–5.

Kutschenreiter-Praszkiewicz, I. (2017). Graph-based decision making in industry. In *Graph Theory-Advanced Algorithms and Applications*. IntechOpen.

Liu, Z., Yao, C., Yu, H., & Wu, T. (2019). Deep reinforcement learning with its application for lung cancer detection in medical Internet of Things. *Future Generation Computer Systems*, *97*, 1–9. https://doi.org/10.1016/j.future.2019.02.068

Loon, R., van. (2018, January 23). *Machine Learning Explained: Understanding Supervised, Unsupervised, and Reinforcement Learning*. Datafloq. https://datafloq.com/read/machine-learning-explained-understanding-learning/

M., B., & P., C. (2016). An Automated Technique using Gaussian Naïve Bayes Classifier to Classify Breast Cancer. *International Journal of Computer Applications*, *148*(6), 16–21. https://doi.org/10.5120/ijca2016911146

MacCallum, P. (2022). *A new treatment for acute myeloid leukemia could prove beneficial for even more people*. https://medicalxpress.com/news/2022-09-treatment-acute-myeloid-leukemia-beneficial.html

Mahmood, N., Shahid, S., Bakhshi, T., Riaz, S., Ghufran, H., & Yaqoob, M. (2020). Identification of significant risks in pediatric acute lymphoblastic leukemia (ALL) through machine learning (ML) approach. *Medical & Biological Engineering & Computing*, *58*(11), 2631–2640. https://doi.org/10.1007/s11517-020-02245-2

Mallick, P. K., Mohapatra, S. K., Chae, G.-S., & Mohanty, M. N. (2020). Convergent learning–based model for leukemia classification from gene expression. *Personal and Ubiquitous Computing*. https://doi.org/10.1007/s00779-020-01467-3

McKenzie, S. B. (2005). Advances in understanding the biology and genetics of acute myelocytic leukemia. *Clinical Laboratory Science: Journal of the American Society for Medical Technology*, *18*(1), 28–37.

MedicalStay Group. (2014). *Leukaemia – BONE MARROW*. https://www.medicalstaygroup.com/leukaemia-bone-marrow/

Mejdoub, M., & Ben Amar, C. (2013). Classification improvement of local feature vectors over the KNN algorithm. *Multimedia Tools and Applications*, *64*(1), 197–218. https://doi.org/10.1007/s11042-011-0900-4

Mignogna, M. D., Fedele, S., & Russo, L. L. (2004). The World Cancer Report and the burden of oral cancer: *European Journal of Cancer Prevention*, 139–142. https://doi.org/10.1097/00008469-200404000-00008

Miyoshi, I., Limited, F., Miwa, S., Inoue, K., & Kondo, M. (2018). *Run-Time DFS/DCT Optimization for Power-Constrained HPC Systems*. 3.

Mostavi, M., Chiu, Y.-C., Huang, Y., & Chen, Y. (2020). Convolutional neural network models for cancer type prediction based on gene expression. *BMC Medical Genomics*, *13*(S5), 44. https://doi.org/10.1186/s12920-020-0677-2

Narayanan, S., & Shami, P. J. (2012). Treatment of acute lymphoblastic leukemia in adults. *Critical Reviews in Oncology/Hematology*, *81*(1), 94–102. https://doi.org/10.1016/j.critrevonc.2011.01.014

Newman, W. M., & Sproull, R. F. (1974). An approach to graphics system design. *Proceedings of the IEEE*, *62*(4), 471–483. https://doi.org/10.1109/PROC.1974.9451

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. (2011). Multimodal Deep Learning. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011* (p. 696). Omnipress.

O'Brien, M. M., Seif, A. E., & Hunger, S. P. (2018). Acute lymphoblastic leukemia in children. In *Wintrobe's Clinical Hematology: Fourteenth Edition* (pp. 4939–5015). Wolters Kluwer Health Pharma Solutions (Europe) Ltd. https://doi.org/10.1056/nejmra1400972

Okun, O. (2011). *Feature selection and ensemble methods for bioinformatics: Algorithmic classification and implementations*. Medical Information Science Reference.

Olanow, C. W., Watts, R. L., & Koller, W. C. (2001). An algorithm (decision tree) for the management of Parkinson's disease (2001): Treatment. *Neurology*, *56*(Supplement 5), S1–S88. https://doi.org/10.1212/WNL.56.suppl_5.S1

Oostindjer, M., Alexander, J., Amdam, G. V., Andersen, G., Bryan, N. S., Chen, D., Corpet, D. E., De Smet, S., Dragsted, L. O., Haug, A., Karlsson, A. H., Kleter, G., de Kok, T. M., Kulseng, B., Milkowski, A. L., Martin, R. J., Pajari, A.-M., Paulsen, J. E., Pickova, J., … Egelandsdal, B. (2014). The role of red and processed meat in colorectal cancer development: A perspective. *Meat Science*, *97*(4), 583–596. https://doi.org/10.1016/j.meatsci.2014.02.011

Özcan ŞİMşek, N. Ö., Özgür, A., & Gürgen, F. (2019). Statistical representation models for mutation information within genomic data. *BMC Bioinformatics*, *20*(1), 324. https://doi.org/10.1186/s12859-019-2868-4

Pan, L., Liu, G., Lin, F., Zhong, S., Xia, H., Sun, X., & Liang, H. (2017). Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia. *Scientific Reports*, *7*(1), 7402. https://doi.org/10.1038/s41598-017-07408-0

Pandya, R., & Pandya, J. (2015). C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning. *International Journal of Computer Applications*, *117*(16), 18–21.

Parmar, C., Grossmann, P., Bussink, J., Lambin, P., & Aerts, H. J. W. L. (2015). Machine Learning methods for Quantitative Radiomic Biomarkers. *Scientific Reports*, *5*(1), 13087. https://doi.org/10.1038/srep13087

Patel, R. (2021). *Predicting invasive ductal carcinoma using a Reinforcement Sample Learning Strategy using Deep Learning* (arXiv:2105.12564). arXiv. http://arxiv.org/abs/2105.12564

Pui, C.-H. (2006). Treatment of Acute Lymphoblastic Leukemia. *N Engl j Med*, 13.

Puzanov, A., Zhang, S., & Cohen, K. (2020). Deep reinforcement one-shot learning for artificially intelligent classification in expert aided systems. *Engineering Applications of Artificial Intelligence*, *91*, 103589. https://doi.org/10.1016/j.engappai.2020.103589

Rahman, S. I., Jadoon, M., Ali, S., Khattak, H., & Huang, J. (2021). Efficient Segmentation of Lymphoblast in Acute Lymphocytic Leukemia. *Scientific Programming*, *2021*, 1–7. https://doi.org/10.1155/2021/7488025

Rai, K. R., Holland, J. F., Glidewell, O. J., Weinberg, V., Brunner, K., Obrecht, J. P., Preisler, H. D., Nawabi, I. W., Prager, D., Carey, R. W., Cooper, M. R., Haurani, F., Hutchison, J. L., Silver, R. T., Falkson, G., Wiernik, P., Hoagland, H. C., Bloomfield, C. D., James, G. W., … Kaan, S. K. (1981). Treatment of acute myelocytic leukemia: A study by cancer and leukemia group B. *Blood*, *58*(6), 1203–1212.

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. *New England Journal of Medicine*, *380*(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

Ratley, A., Minj, J., & Patre, P. (2020). Leukemia Disease Detection and Classification Using Machine Learning Approaches: A Review. *2020 First International Conference on*

Power, Control and Computing Technologies (ICPC2T), 161–165. https://doi.org/10.1109/ICPC2T48082.2020.9071471

Saini, I., Singh, D., & Khosla, A. (2013). QRS detection using K-Nearest Neighbor algorithm (KNN) and evaluation on standard ECG databases. *Journal of Advanced Research*, *4*(4), 331–344. https://doi.org/10.1016/j.jare.2012.05.007

Sammut, C., & Webb, G. I. (Eds.). (2010). *Encyclopedia of Machine Learning*. Springer US. https://doi.org/10.1007/978-0-387-30164-8

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, *270*(5235), 467–470. https://doi.org/10.1126/science.270.5235.467

Sharma, A., & Rani, R. (2017). An Optimized Framework for Cancer Classification Using Deep Learning and Genetic Algorithm. *Journal of Medical Imaging and Health Informatics*, *7*(8), 1851–1856. https://doi.org/10.1166/jmihi.2017.2266

Simin, A. T., Mohsen Ghorabi Baygi, S., & Noori, A. (2020). Cancer Diagnosis Based on Combination of Artificial Neural Networks and Reinforcement Learning. *2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, 1–4. https://doi.org/10.1109/ICSPIS51611.2020.9349530

Stewart, B. W., & Wild, C. P. (Eds.). (2014). *World cancer report 2014*. International Agency for Research on Cancer.

Sun, Y., Zhu, S., Ma, K., Liu, W., Yue, Y., Hu, G., Lu, H., & Chen, W. (2019). Identification of 12 cancer types through genome deep learning. *Scientific Reports*, *9*(1), 17256. https://doi.org/10.1038/s41598-019-53989-3

Sutton, R. S., & Barto, A. G. (2015). *Reinforcement Learning: An Introduction*. 398.

Tang, B., Pan, Z., Yin, K., & Khateeb, A. (2019). Recent Advances of Deep Learning in Bioinformatics and Computational Biology. *Frontiers in Genetics*, *10*. https://www.frontiersin.org/articles/10.3389/fgene.2019.00214

Taqman. (2009). *Taqman Genes Expression arrays*. https://www.thermofisher.com/sa/en/home/life-science/pcr/real-time-pcr/real-time-pcr-assays.html

Tarek, S., Abd Elwahab, R., & Shoman, M. (2017). Gene expression based cancer classification. *Egyptian Informatics Journal*, *18*(3), 151–159. https://doi.org/10.1016/j.eij.2016.12.001

Terwilliger, T., & Abdul-Hay, M. (2017). Acute lymphoblastic leukemia: A comprehensive review and 2017 update. *Blood Cancer Journal*, *7*(6), e577–e577. https://doi.org/10.1038/bcj.2017.53

Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). Review The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Współczesna Onkologia*, *1A*, 68–77. https://doi.org/10.5114/wo.2014.47136

Verda, D., Parodi, S., Ferrari, E., & Muselli, M. (2019). Analyzing gene expression data for pediatric and adult cancer diagnosis using logic learning machine and standard supervised methods. *BMC Bioinformatics*, *20*(S9), 390. https://doi.org/10.1186/s12859-019-2953-8

Wang, K., An, Y., Zhou, J., Long, Y., & Chen, X. (2022). A novel Multi-Level feature selection method for radiomics. *Alexandria Engineering Journal*, S1110016822007268. https://doi.org/10.1016/j.aej.2022.10.069

Wen, J., Xu, Y., Li, Z., Ma, Z., & Xu, Y. (2018). Inter-class sparsity based discriminative least square regression. *Neural Networks*, *102*, 36–47. https://doi.org/10.1016/j.neunet.2018.02.002

Wright, R. E. (1995). *Logistic regression.*

Xu, B., Liu, J., Hou, X., Liu, B., Garibaldi, J., Ellis, I. O., Green, A., Shen, L., & Qiu, G. (2019). *Look, Investigate, and Classify: A Deep Hybrid Attention Method for Breast Cancer Classification* (arXiv:1902.10946). arXiv. http://arxiv.org/abs/1902.10946

Zahurak, M., Parmigiani, G., Yu, W., Scharpf, R. B., Berman, D., Schaeffer, E., Shabbeer, S., & Cope, L. (2007). Pre-processing Agilent microarray data. *BMC Bioinformatics*, *8*(1), 142. https://doi.org/10.1186/1471-2105-8-142

Zhang, X.-D. (2020). *A Matrix Algebra Approach to Artificial Intelligence*. Springer Singapore. https://doi.org/10.1007/978-981-15-2770-8

Zhang, Y. (2019). Classification and Diagnosis of Thyroid Carcinoma Using Reinforcement Residual Network with Visual Attention Mechanisms in Ultrasound Images. *Journal of Medical Systems*, *43*(11), 323. https://doi.org/10.1007/s10916-019-1448-5

Zhao, J., Zhang, M., Zhou, Z., Chu, J., & Cao, F. (2017). Automatic detection and classification of leukocytes using convolutional neural networks. *Medical & Biological Engineering & Computing*, *55*(8), 1287–1301. https://doi.org/10.1007/s11517-016-1590-x

Zhao, L., Chen, Y., & Schaffner, D. W. (2001). Comparison of Logistic Regression and Linear Regression in Modeling Percentage Data. *Applied and Environmental Microbiology*, *67*(5), 2129–2135. https://doi.org/10.1128/AEM.67.5.2129-2135.2001

BIOGRAPHY

Zaid Mohammed Ibrahim IBRAHIM. He received his B.Sc. degree in Electronics and Communications Engineering Department from Al-Nahrain University in 2014. He has been working as a Data Analyst since 2018. His research areas are Neural Networks, IoT, Big Data, Digital Signal Processing, and Computer Vision.