Machine learning-based cancer prediction using large scale clinical data

by

Oladotun Osisami

M.S., University of Louisiana at Lafayette, 2010

A REPORT

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Chemical Engineering
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2024

Approved by:

Major Professor
Davood B. Pourkargar

Copyright

Abstract

This study delves into the urgent challenge of cancer prediction by utilizing machine learning techniques on extensive clinical data. Conventional diagnostic techniques frequently result in delays and low sensitivity, underscoring the need for novel strategies. This work focuses on miRNA expression patterns from the Genomic Data Commons (GDC) by utilizing machine learning (ML) on large-scale clinical data. It builds a neural network model using the Keras Sequential API, investigates five machine learning approaches, and applies feature selection strategies to improve model interpretability. The research aims to provide insights into improving early cancer detection and risk assessment. Despite inherent limitations such as data quality variability and computational constraints, the study aims to rigorously examine ML methodologies in cancer prediction, with implications for future research and practical applications.

Utilizing Python and a Jupyter notebook, this study gathered miRNA expression data from the Genomic Data Commons (GDC) via its API, ensuring data quality through preprocessing techniques like cleaning and normalization. Feature selection based on mutual information scores was then applied to enhance model interpretability and performance. Subsequently, five machine learning methods (k-nearest neighbors, random forest, logistic regression, support vector machine, and gradient boosting) were employed for cancer type classification, alongside a neural network model crafted with the Keras Sequential API for multi-class cancer classification. Performance evaluation metrics, including accuracy, precision, recall, and F1-score, were computed to assess model discriminative capabilities.

This study investigates cancer classification via machine learning on genomic data, revealing diverse gene expression profiles across different cancer types. Preprocessing and feature selection

resulted in a dataset of 2,507 samples and 1,881 features, addressing class imbalances. Support Vector Machine achieved notable performance, with high accuracy (99.04%), precision (96%), recall (98%), and F1-score (98.98%) across all cancer types. Random Forest demonstrated precision scores of 98% and an F1-score of 98.98%. Logistic Regression displayed robust performance across metrics, while Gradient Boosting showed strong accuracy and precision. K-Nearest Neighbors exhibited moderate accuracy and precision. Neural Network performed consistently well across metrics, with rapid convergence and high accuracy on unseen data. Confusion matrices and ROC curves validated accurate predictions, highlighting the potential of machine learning in precise cancer classification and early intervention.

In conclusion, the five classifiers demonstrated robustness in accurately distinguishing between different cancer types with minimal false positives with Support vector machine standing out for its outstanding performance, with an accuracy score of 0.9904, reaffirming its efficacy in cancer classification tasks

Table of Contents

List of Figures

List of Tables

Acknowledgement

I would like to extend my deepest gratitude to my wife, Paula Osisami, and my three children for their unwavering support, patience, and understanding throughout this journey. Their love and encouragement have been my source of strength, enabling me to pursue and complete this MSc. program. I am profoundly grateful for their presence in my life.

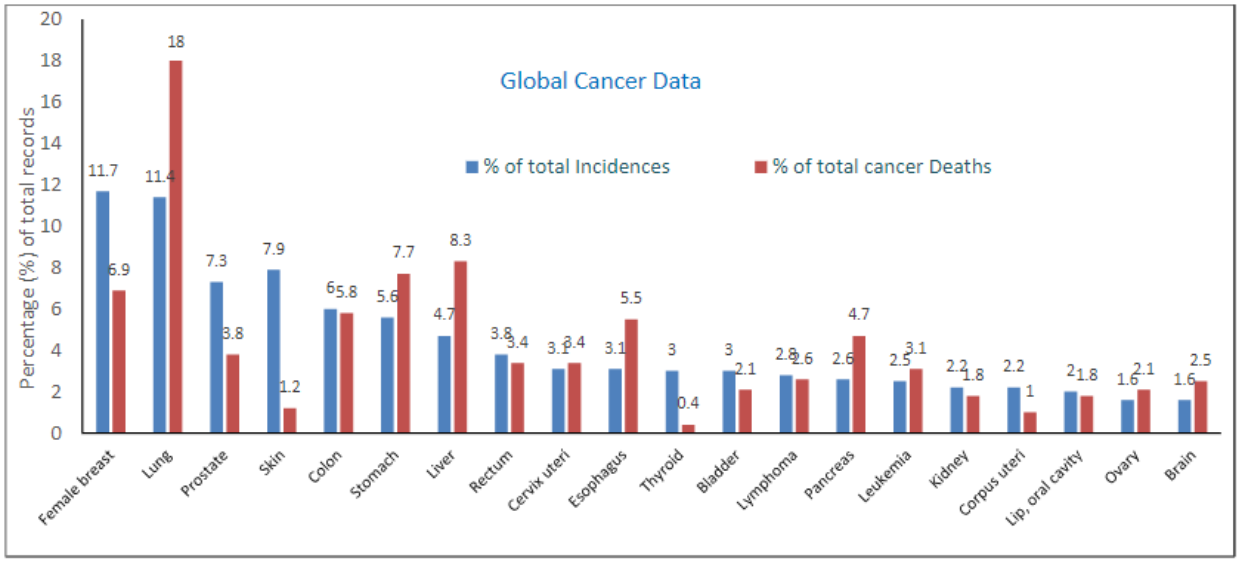
Introduction

1.1 Background Study

Cancer continues to be one of the most pressing public health challenges worldwide, with its prevalence and impact on individuals and societies escalating each year. In the medical world, cancer first appeared in the 1600s and is defined by the growth of aberrant cells that can invade or spread to other parts of the body (Abraham *et al.* 2009). Through a process called cancer metastasis, this unchecked cellular development starts in one area of the human body and spreads to other parts (Seyfried and Huysentruyt 2013; Liotta 1992). Usually, these cancerous cells are categorized as benign or malignant. Malignant cells metastasis and present a more significant hazard because of their destructive nature, whereas benign cells stay localized and do not spread. Cancer has a long and expensive treatment regimen, primarily because of its high death rates and propensity for recurrence. An estimated 1.7 million new cases and 0.6 million deaths were attributed to cancer in the US in 2019 (Huang *et al.* 2020). According to stated data, there have been around 19.3 million (19,300,000) newly diagnosed cancer cases recently, which would result in about 10 million fatalities in 2020 (Chhikara and Parang 2023; Ferlay *et al.* 2021). (Figure 1).

Traditional methods of cancer diagnosis rely heavily on invasive procedures such as biopsies and imaging tests, often leading to delays in detection and diagnosis (Schiffman *et al.* 2015). Moreover, these methods may not be sensitive enough to detect early-stage tumors or identify high-risk individuals who would benefit from preventive interventions. As the number of cancer cases and deaths rises, newer surgical methods like robotic surgery and laparoscopy, along with tumor adjuvant therapy and other cutting-edge technologies are becoming increasingly necessary to improve survival rates and lower local recurrence rates (Shinji *et al.* 2022; Akagi and Inomata 2020).

Figure 1. Percentage Representation of Total Cancer Incidences (19.3 million) and Total Deaths (10 million). Data source: Chhikara and Parang (2023)



Adapted with permission from "Chemical Biology Letters," vol. 10, no. 1 (2023): 451.

As chemical engineers, irrespective of whether it involves discovering new pharmaceutical ingredients for combating diseases or enhancing process efficiencies to comply with stricter environmental laws, it is necessary to possess the ability to predict outcomes of certain events. These events range from managing heat supply in reactors to determining reaction rates and selectivity. Theoretical models, honed over centuries, aid in predictions, but many require substantial computational resources for numerical solutions, as analytical solutions aren't feasible for realistic systems(Dobbelaere *et al.* 2021).

Machine Learning, a subset of Artificial Intelligence, aligns the task of learning from datasamples with the broader concept of inference (Dunjko and Briegel 2018; Alzubi *et al.* 2018). The two primary stages of the learning process are usually (i) using an existing dataset to estimate unknown dependencies inside a system and (ii) using these estimated dependencies to predict new system outputs (Kourou *et al.* 2015). Machine learning is becoming an exciting field with many biomedical research applications. A given collection of biological samples entails applying various approaches and algorithms to navigate an n-dimensional space in search of appropriate generalizations (Niknejad and Petrovic 2013; Kourou *et al.* 2015). In recent years, significant advancements in machine learning algorithms and the availability of large-scale clinical datasets have fueled rapid progress in cancer prediction and risk stratification. Supervised learning algorithms, such as logistic regression, decision trees, random forests, support vector machines, and deep learning models (e.g., convolutional neural networks), have been widely employed for predictive modeling tasks in oncology (Murali *et al.* 2020; Erdem and Bozkurt 2021).

Numerous studies have demonstrated the feasibility and effectiveness of machine learning-based approaches for cancer prediction using large-scale clinical data. Researchers have developed predictive models for various cancer types, including breast cancer (CHAUDHURY *et al.* 2021),

lung cancer (Cengil and Cinar 2018), colorectal cancer (Bychkov *et al.* 2018), prostate cancer (Singh *et al.* 2023), and others. These models have shown promising results in accuracy, sensitivity, specificity, and clinical utility, paving the way for their integration into routine clinical practice.

Applying machine learning algorithms to clinical data offers a novel and potentially transformative approach to cancer prediction and risk stratification. By leveraging large-scale datasets containing diverse patient information, including demographic, genetic, clinical, and imaging data, machine learning models can identify complex patterns and relationships that may not be apparent through conventional analysis methods. This enables the development of predictive models capable of identifying individuals at increased risk of developing cancer, facilitating early intervention and personalized treatment strategies.

1.2 Research Objectives

This research aims to develop and evaluate machine learning models for cancer prediction using large-scale clinical data, focusing on miRNA expression profiles obtained from the Genomic Data Commons (GDC). Specifically, the study objectives are:

i. Collect miRNA expression data from Genomic Data Commons (GDC) and preprocess it for analysis.
ii. Explore five machine learning methods and optimize dataset characteristics for classification.
iii. Select informative features using Mutual Information scores to enhance model interpretability.
iv. Train six classifiers independently and evaluate their performance for cancer classification.

v. Build and assess the performance of a neural network model for cancer classification using Keras.

1.3 Structure of the Report

The remainder of this thesis is organized as follows:

Chapter 2 comprehensively reviews the literature on cancer prediction, machine learning techniques, and clinical data integration in predictive modeling.

Chapter 3 outlines the methodology employed in this study, including data collection, preprocessing, feature selection, model development, and evaluation metrics.

Chapter 4 presents the experiments' results, including model performance metrics, comparative analyses, and interpretation of findings.

Chapter 5 discusses the implications of the study findings, potential limitations, and avenues for future research.

Finally, Chapter 6 summarizes the study's key findings and offers conclusions and recommendations for further research and practical applications.

1.4 Scope and Limitations

1.4.1 Scope

The scope of this study encompasses developing and evaluating machine learning models for cancer prediction using miRNA expression data obtained from the Genomic Data Commons (GDC). The study explores five well-known machine learning methods: k-nearest neighbors (KNN), gradient boosting(GB), support vector machine (SVM), random forest(RF), and logistic

regression(LR). Additionally, the scope includes feature selection techniques based on Mutual information (MI) scores to enhance model interpretability and classification performance. The methodology also involves constructing and evaluating a neural network model(NN) using the Keras Sequential API for cancer classification. The study aims to contribute to the field of cancer research by providing insights into the utility of machine learning techniques in improving early detection and risk assessment.

1.4.2 Limitations

This study, while ambitious in its scope, is subject to several inherent limitations that may temper the breadth and applicability of its findings. Among these limitations is the potential variability in the availability and quality of miRNA expression data obtained from the Genomic Data Commons (GDC), which may notably influence the reliability and generalizability of the machine learning models developed. Moreover, the efficacy of the models is contingent upon the unique characteristics of the dataset and the specific cancer types under consideration, thereby necessitating caution in extrapolating findings to broader contexts. Additionally, the feature selection process may inadvertently overlook pertinent features germane to cancer prediction, warranting further investigation into alternative feature selection methodologies or the integration of domain-specific knowledge. Notably, despite efforts to bolster model interpretability through visualization techniques, specific machine learning models, particularly NN, may exhibit diminished interpretability due to their inherent complexity. Lastly, the computational demands entailed in training and evaluating multiple machine learning models, particularly NN, may impose constraints on the scalability and complexity of the analyses conducted in this study, thus warranting judicious allocation of computational resources.

In recognition of these constraints, this study endeavors to provide a transparent and methodologically rigorous examination of the application of machine learning methodologies in cancer prediction while remaining cognizant of the attendant limitations that may impact the robustness and generalizability of its findings.

1.5 Definition of Terms

miRNA Expression Data: Quantitative measurements of microRNA levels in biological samples crucial for understanding gene regulation and identifying biomarkers in cancer research.

Genomic Data Commons (GDC): A centralized repository managed by the National Cancer Institute (NCI) that hosts diverse genomic and clinical datasets, including miRNA expression data, to facilitate collaboration in cancer genomics research.

Machine Learning Methods: Algorithms enabling computers to learn from data and make predictions without explicit programming.

Feature Selection: Identifying pertinent dataset features to enhance model performance and interpretability.

Neural Network Model: Computational models inspired by the structure of biological neural networks.

Literature Review

2.1 Cancer

Cancer, a formidable challenge in health, comprises a diverse array of diseases marked by the uncontrolled growth and spread of abnormal cells within the body (Nenclares and Harrington 2020). The complex characteristics of this disease lead to different types of cancer, each with distinct difficulties in diagnosis, management, and prognosis. Over 20 million additional instances of cancer are expected annually by 2025, according to global demographic trends, which point to an increasing incidence of the disease in the future decades (Valery *et al.* 2018).

Based on GLOBOCAN data from 2012, there were 8.2 million cancer-related deaths and about 14.1 million new cases of cancer (Ferlay *et al.* 2013). Prostate, lung, bronchus, colon and rectum, and urinary bladder cancers were the most common cancer types in men. The breast, lung and bronchus, colon and rectum, uterine corpus, and thyroid were the areas in which women experienced the highest incidences. The aforementioned data indicates that a considerable proportion of cancer cases in men and women are related to prostate and breast cancer, respectively (Ward *et al.* 2019). The most common malignancies in children were those of the blood, brain, and lymph nodes (Miller *et al.* 2020). Incidence and mortality from cancer are still caused mainly by lung cancer worldwide (Thandra *et al.* 2021). Worldwide, there were projected to be 19.3 million new cases of cancer in 2020, along with roughly 10 million cancer-related deaths (Li 2022). The World Health Organization highlights the critical need for novel strategies to comprehend and treat cancer, pointing out that it is a leading cause of illness and mortality worldwide (W.H.O., 2018).

2.2 Cancer Diagnosis

Cancer diagnosis is a complex process in modern oncology, continually advancing with improvements in medical technology, diagnostic techniques, and understanding of cancer biology.

During the preliminary stage, clinical assessments conducted by healthcare professionals play a vital role in identifying symptoms and potential signs of malignancy (Sawicki *et al.* 2021). Subsequently, imaging technologies like X-rays, CT scans, MRI, and PET scans provide detailed images of tumors (Mankoff 2007). Biopsy procedures, where tissue samples are extracted and analyzed, confirm cancer diagnosis by identifying cancer cells and determining tumor characteristics (De Mattos-Arruda *et al.* 2015). Molecular pathology and genomic profiling have revolutionized cancer diagnosis by allowing the examination of tumors at a genetic level, leading to personalized treatment approaches (Roychowdhury *et al.* 2011). Via the use of liquid biopsy techniques and blood-based biomarkers such as circulating tumor cells (CTCs) and circulating tumor DNA (ctDNA), diagnostic capacities can be improved by non-invasively tracking tumor activity in real-time (Soda 2021).

2.3 Early Detection of Cancer

Early cancer detection greatly increases the chances of survival. On the other hand, the majority of malignancies are discovered at an advanced stage (Llewellyn *et al.* 2018). Improving cancer survival rates requires early identification of the disease. The goal of early detection is to find signs of cancer or precancerous changes as soon as possible, allowing for prompt treatment to increase survival chances and reduce complications. This process involves identifying cell abnormalities that may lead to cancer, including cancer itself and early changes that could develop into cancer (Blandin Knight *et al.* 2017). Screening is a component of early detection that involves testing people who do not exhibit symptoms (Raffle and Gray 2019). For example, identifying minimal residual disease and cancer recurrence is directly related to these principles and other facets of cancer care. The ultimate objective is to identify and treat possible health problems before they become life-threatening or seriously impair an individual's quality of life.

Early cancer detection is a multi-stage process that begins with regular cell activity and advances to cancer development (Rezaeipanah and Ahmadi 2022). This entails recognizing the malignancy as well as any precursor changes that can eventually result in cancer. Testing individuals who do not exhibit symptoms is known as screening, one facet of early detection initiatives. The principles of early diagnosis are intimately related to various facets of cancer care, including the identification of minimal residual illness and cancer recurrence (Tran *et al.* 2021). The ultimate goal is to identify and manage possible health problems before they cause severe illness or death within an individual's expected lifespan.

While early identification boosts survival in all groups, around 70% of cancer deaths happen in low- and middle-income nations (Crosby *et al.* 2022), frequently as a result of delayed diagnosis. For example, from the 1970s to 2011, Black sub-Saharan African women's rate of late-stage breast cancer detection consistently maintained over 60%, but in the US, Black women's incidence of late diagnosis dropped from roughly 60% to 32% during the same period (Jedy-Agba *et al.* 2016). Certain cancers with practical early detection tests, like cervical cancer, exhibit significantly higher mortality rates in low human development index (HDI) countries compared to high HDI countries (19.8 versus 3.1 deaths per 100,000, respectively), while other cancers without early detection tests show less variation (e.g., stomach cancer, 5.0 versus 4.0 deaths per 100,000, respectively) (Crosby *et al.* 2022).

A significant equity issue is raised by the worldwide dilemma of late-stage cancer identification, which is made worse in environments with limited resources (Haier and Schaefers 2022). When a patient receives a later-stage cancer diagnosis, they may forfeit the chance for a curative intervention, and costly systemic treatments frequently result in worse outcomes and severe side effects. Significant improvements in patient outcomes could result from additional research

focused on applying early detection techniques to different forms of cancer.

2.4 Importance of cancer detection

It is impossible to overstate the significance of early cancer detection, as it significantly influences patient outcomes, treatment approaches, and overall survival rates. Early cancer detection makes it possible to identify the disease earlier when it is more contained and potentially treatable. At this point, different therapeutic interventions, such as surgery, targeted therapies, and chemotherapy, have a higher chance of being successful (Lee *et al.* 2018). Removing tumors surgically or administering targeted treatments in an early stage greatly increases the likelihood of successful treatment.

Early identification of cancer often enables more targeted and less aggressive treatment options (Schiffman *et al.* 2015). For instance, cancer may require more involved and invasive therapies if it is discovered in its later stages, which could increase side effects and lower the patient's quality of life. Early detection can lessen the need for these intense interventions, easing the physical and psychological toll that therapy takes on patients.

Early cancer detection has financial advantages for healthcare systems and improves patient outcomes. According to Mariotto et al. (2011), treating cancer in its early stages is less expensive than handling instances that have progressed. The latter entails more significant costs associated with extended hospital stays and costly procedures. Healthcare systems that want to deploy resources as efficiently as possible must prioritize this cost-effectiveness.

The direct link between early identification and higher survival rates is possibly the most robust case in favor of it. Early cancer detection increases the likelihood that it will be successfully treated, extending the affected person's life expectancy (Allemani *et al.* 2018). The significant

influence that early identification can have on the overall prognosis of cancer patients is highlighted by the disparity in survival rates between early and late-stage diagnoses.

It also significantly enhances the biological well-being of cancer survivors. Early intervention frequently involves less invasive treatments to reduce physiological stress and expedite recovery (Burback *et al.* 2024). People can thereby maintain their best possible physiological state and general well-being throughout and after treatment.

Programs for systematic early detection have the potential to improve population health generally. These programs can potentially lower the total burden of cancer and related healthcare expenditures by detecting and treating cases of the illness at an early stage (Cancer Research UK, 2020) (UK 2020). Initiatives for population-level screening for common malignancies, including colorectal, cervical, and breast cancer, have shown the value and viability of early detection techniques (Dare *et al.* 2021).

2.5 Challenges Faced with Early Diagnosis

According to Dagogo-Jack et.al (2018), there are still ongoing issues in the diagnosis of cancer because the disease is heterogeneous, there are gaps in early detection techniques, and there are differences in the accessibility of diagnostic assistance. As a significant advance in accuracy, especially in radiological imaging and pathology, the integration of artificial intelligence (AI) and machine learning (ML) in cancer detection constitutes a breakthrough (Topol 2019). As advancements continue, the focus is on improving existing procedures, investigating novel diagnostic approaches, and tackling the complex problems associated with providing a cancer diagnosis to those affected.

2.6 Machine Learning

Machine learning has revolutionized medical diagnosis, particularly in cancer, where accurate tumor diagnosis is crucial (Iqbal *et al.* 2021). Classifying tumors can be challenging because lesions can range from benign to malignant. Using machine learning algorithms to analyze medical imaging data is an innovative way to increase the efficiency and accuracy of malignant tumor classification (Erickson *et al.* 2017).

Cancerous tumors can vary widely in size, form, and properties depending on their origin. Tumors are commonly detected by imaging techniques such as magnetic resonance imaging (MRI), computed tomography (CT), and ultrasound (Kong *et al.* 2019). Without a thorough evaluation, it can be challenging to determine the precise type of tumor and distinguish between benign and malignant growths.

Machine learning transforms traditional rule-based techniques for classifying cancer tumors into automated, data-driven procedures (Sarker 2021). These systems, equipped with sophisticated algorithms, can process extensive image collections, recognize intricate patterns, and provide quick, precise predictions. The primary objective is to support medical practitioners in making well-informed decisions that may lead to early interventions, customized treatment plans, and improved patient outcomes for all types of cancer.

2.6.1 Supervised learning algorithm for tumor classification

A key component of machine learning is supervised learning (SL), where the objective is to acquire the ability to precisely map input data to match output labels (Abdulqader *et al.* 2020). Each instance of input data in this learning process is linked with a predefined output label through the use of labeled training data. The algorithm analyzes patterns and relationships in the data by

analyzing these annotated samples. The ultimate goal is to create a model capable of precisely predicting outputs for brand-new, untested input data. Figure 2.1 shows a flowchart for supervised learning.

Support Vector Machine (SVM)

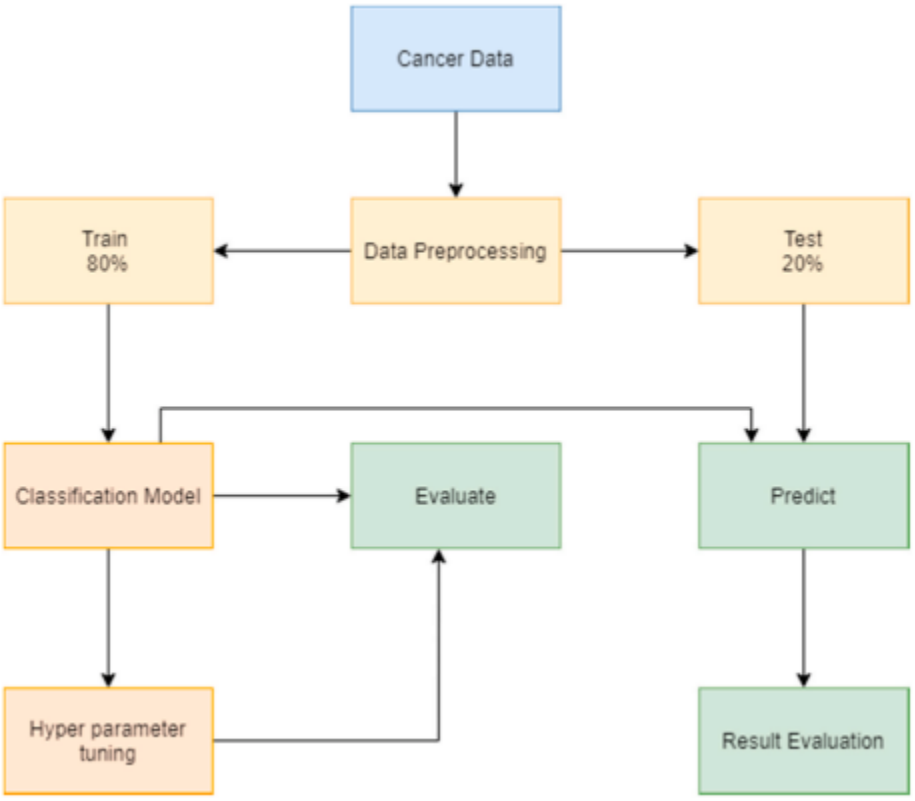
A reliable and adaptable method used to address a wide range of classification issues is the SVM (Awad *et al.* 2015). Its fundamental idea is to find the best boundary—a hyperplane—that maximizes the distance between the dataset's various categories. This hyperplane is positioned to be equally distant from the nearest data points in each category to create a margin depicting the separation degree. A more considerable margin usually indicates higher categorization accuracy. Determining the hyperplane depends heavily on support vectors, essential data points on the margin's edge. The efficacy of SVM is not limited to a single dataset type; it may also be applied to datasets with complex patterns that are difficult to divide using a straight line (Campbell and Ying 2022; Ghosh *et al.* 2019; Gaye *et al.* 2021). It uses many methods, such as linear approaches and the radial basis function (RBF), to identify the best hyperplane for accurately classifying the data. SVM is appropriate for both small- and large-scale applications since it can handle datasets of different sizes and performs well in high-dimensional environments (Ghosh *et al.* 2019).

The SVM optimizes the separation between the separating hyperplane and the support vectors to offer the best possible separability between classes (de Boves Harrington 2015). The mathematical definition of the linear separating hyperplane in the feature space is as follows:

$$y = l^T x_i + d$$

where x_i is a feature vector of a binary class, d is an intercept, and l is the norm vector of the

Figure 2. Supervised Learning Flowchart



separating hyperplane for the two classes. The cost function of l is expressed as:

$$S(l,k) = \frac{1}{2}||l||^2 + C\sum_{i=1}^{E} k_i$$

Subject to:

$$y_i \times (l^T x_i + d) \geq (1 - k_i) \qquad k_i \geq 0$$

where C is the regularization parameter, k_i indicates the classification errors, E represents the number of misclassified samples, and y_i denotes the class label.

Logistic Regression (LR)

LR is a machine learning technique mainly applied to binary classification issues, meaning the solution is either yes or no (Strickland 2017). It uses a mathematical model designed to forecast things based on input data. Because of its adaptability, LR can forecast whether an observation falls into a specific category, making it suitable for handling progressively complicated circumstances. The benefits of LR include low preprocessing requirements for input data, ease of training, computing efficiency, and simplicity (Awad *et al.* 2023). However, its inability to solve more complicated problems and occasional tendency to focus too much on details can be a disadvantage.

Decision Tree Algorithm

The decision tree (DT) is one kind of categorization algorithm. It grows the tree using a divide-and-conquer tactic (Priyanka and Kumar 2020). The data instances are connected by a variety of characteristics. Nodes and leaves comprise a decision tree. A decision tree's leaves display the kinds of events that satisfy the criteria, while nodes represent checks on feature values. The answer is either "true" or "false". Rules can be constructed from the root node to the leaf, and the nodes

met on the way are required for the rule to predict the class at the leaf (Song and Ying 2015). Trees must be pruned to remove unnecessary preconditions and redundancy.

Random Forest (RF)

Building a strong "forest" out of several decision trees is the basis of the RF, a supervised learning method that may be used for both regression and classification tasks (Quatrini *et al.* 2020). RF builds decision trees from various datasets, combines their predictions, and uses a voting process to choose the best option. This process is similar to how a forest becomes stronger with more trees. RF reduces overfitting by combining results collaboratively, outperforming a single decision tree. RF creates a varied forest by combining several decision trees and using feature randomness and bagging to construct a classification rule set (Dhilsath and Samuel 2021). The combined predictions of all the trees provide a more precise and complex outcome than any one tree could have done on its own.

K-Nearest Neighbors (KNN)

A straightforward machine learning technique for classifying data is the KNN algorithm. To identify the most similar data points, it compares newly collected data with existing ones (Shokrzade *et al.* 2021). Based on the classification of the most comparable points, the category of the new data point is then guessed. KNN is renowned for being straightforward and for doing well with noisy data. For precise forecasts, it is crucial to determine the appropriate number ("k") of similar points to consider. KNN can process various kinds of data; however, it might not function effectively in massive datasets or where the data is poorly organized (Ali *et al.* 2019). KNN is a helpful machine learning tool overall, particularly for smaller tasks or where simplicity is preferred.

Gradient Boosting

Known for its accuracy in predicting outcomes, GB is a potent machine-learning approach (Zhang and Haghani 2015). It functions through a sequential combination of several basic models, such as decision trees. All subsequent models fix the errors of their predecessors. GB builds a robust prediction model by assigning greater weight to models that accomplish well (Bentéjac *et al.* 2021). Tasks like category, quantity, or ranking prediction are done with it. GB has become even more efficient with the development of variants like AdaBoost and XGBoost (Bentéjac *et al.* 2021; Aziz *et al.* 2020). For many prediction problems, this approach is helpful because it works well with complex data relationships. Since Friedman and his colleagues first proposed the concept of GB in 2000, it has grown to be one of the most effective machine learning techniques (Freeman *et al.* 2016).

2.6.2 Unsupervised learning

Unsupervised learning is a data-driven method that analyzes unlabeled data in isolation from people (Naeem *et al.* 2023). This technique is often used for exploratory tasks, significant trend and structure finding, generative feature extraction, and data cluster identification. Activities including feature learning, density estimation, grouping, dimensionality reduction, anomaly detection, and association rule discovery are all included in unsupervised learning (Malik and Tuckfield 2019). Unlike supervised learning, unsupervised learning does not depend on predetermined correct or wrong answers or outside assistance. Instead, computers find and expose interesting structures in the data on their own. Algorithms for unsupervised learning take input data and use the features it has learned to classify fresh data. This approach is especially beneficial for feature reduction and clustering, which improves our comprehension of complex relationships and patterns in the data (Alelyani *et al.* 2018).

2.6.3 Semi-supervised learning

Semi-supervised learning is a hybrid strategy, amalgamating elements of supervised and unsupervised approaches by leveraging labeled and unlabeled data (Yang *et al.* 2022). The main goal of a semi-supervised learning model is to improve prediction performance beyond what is possible with labeled data alone. This methodology finds valuable applications in machine translation, fraud detection, text categorization, data labeling, and other areas (Duarte and Berton 2023). Semi-supervised learning, as opposed to traditional supervised techniques that depend on labeled data, contains elements of both supervised and unsupervised machine learning approaches and is especially helpful when getting labeled data is a time-consuming procedure.

2.7 Integration of Multi-Modal Data for Cancer Classification

Machine learning is used to classify cancer by capturing different aspects of cancer biology and patient characteristics using numerous data types. To create reliable and accurate machine learning models for cancer diagnosis, prognosis, and treatment, several data categories offer abundant information.

2.7.1 Gene Expressions

Machine learning approaches have significantly impacted gene expression analysis, especially in detecting and categorizing cancer. This is essential for identifying cancer early and comprehending its underlying genetic traits. Genome research has entered a revolutionary phase, transitioning from classic machine learning techniques to deep learning approaches.

Machine learning can effectively predict cancer subtypes in several investigations, such as those conducted by

Chen et al. (2014) and Golub et al. (1999). Darweesh et al. (2021) have shown the prospective performance of decision tree models that use hierarchical techniques and particle swarm

optimization algorithms for breast cancer screening using mammography pictures. In research like Ali et al. (2023), it has been shown that simpler algorithms like SVM and RF work well for classifying microarray breast cancer data. The efficacy of ensemble-based approaches in classifying lung cancer is a noteworthy example of this research area's complexity (Pradhan *et al.* 2023).

Conclusively, investigations highlight the potential of machine learning methods in precisely classifying cancer kinds according to gene expression information, providing paths for enhanced diagnosis and treatment approaches. Ontology-driven machine learning models and empirical assessments of neural network-based approaches for gene expression analysis are two examples of the ongoing research gaps. Machine learning applications in cancer classification can be advanced through targeted analysis of particular cancer types, including ovarian cancer, as Merlin & Sathiaseelan (2021) noted.

2.7.2 Imaging Analysis

The field of cancer classification using machine learning in conjunction with imaging data analysis is developing quickly and has great promise for influencing cancer detection and therapy. Researchers want to reliably diagnose and describe distinct types of cancer by automating the processing of many medical imaging scans, such as CT, MRI, mammography, and PET scans, by utilizing machine learning techniques, including convolutional neural networks (CNNs) and SVMs (Lee and Chen 2015; Zhang and Sejdić 2019). Machine learning algorithms can accurately distinguish malignant and benign tissues since they are trained on annotated datasets to identify discriminative features suggestive of malignant tumors (Saba 2020). Moreover, a thorough understanding of tumor biology and behavior is made possible by integrating data from many

imaging modalities, making accurate tumor segmentation and measurement for treatment planning and monitoring possible.

It has been widely studied how to classify cancer using machine learning on imaging data, particularly in the detection of brain tumors and breast cancer. The need for machine learning for precise histopathology-based breast cancer diagnosis was highlighted by Saxena & Gyanchandani (2020). The potential of deep learning in medical image analysis was brought to light by Litjens et al. (2017). This is important for tasks like picture classification and segmentation in cancer classification. With an emphasis on computational opportunities and problems, Manhas et al. (2021) evaluated automated cancer diagnosis utilizing deep learning and machine learning techniques. While Khan et al. (2023) focused on MRI-based brain tumor image classification using CNN, Darweesh et al. (2021) proposed a hierarchical machine learning approach for early breast cancer diagnostics based on mammography images, highlighting the growing interest in medical image classification for brain tumors.

All of this research demonstrates how crucial machine learning and deep learning methods are for accurately classifying cancer from medical imaging data, particularly for diagnosing breast cancer and brain tumors. Nevertheless, there remains a research gap in the application of machine learning for the classification of cancer using imaging data in additional cancer types, like ovarian, prostate, and lung cancer. To help create reliable and broadly applicable models for cancer detection and classification, future studies may investigate machine-learning approaches for precise cancer classification across a range of imaging modalities and datasets.

2.8 Current state of cancer research

Applying machine learning techniques in cancer prediction research is essential in transforming cancer treatment and diagnosis approaches. Machine learning algorithms encompass supervised

and unsupervised learning methods widely used in cancer prediction, such as risk evaluation, early identification, tumor characterization, therapy response prediction, and survival analysis. Prominent developments in risk assessment include thoroughly examining extensive epidemiological databases, which forecast an individual's propensity for cancer by considering various multifactorial factors (Kourou et al., 2015). Liu et al. (2019) have proven that machine learning techniques, specifically those that employ CNNs, provide encouraging results when it comes to automating lesion detection and categorization in medical imaging data. This enhances diagnostic accuracy.

Furthermore, machine learning techniques significantly advance the characterization of tumors and their classification by using genomic and clinical data to predict therapy response, identify molecular subtypes of tumors, and tailor effective treatment plans. Coudray et al. (2018) provide an example of how deep learning techniques enhance treatment results for cancer patients by enabling the development of predictive models based on gene expression profiles. These developments highlight how machine learning in cancer prediction has the potential to revolutionize the field by offering previously unattainable insights into the biology of cancer and the mechanisms underlying treatment response.

Machine learning-based cancer prediction research currently includes many studies using different data modalities, and approaches focused on improving our knowledge of cancer prognosis, risk assessment, and treatment results. Together, these studies demonstrate the multidisciplinary nature of cancer prediction research, from the extensive cancer statistics provided by Siegel et al. (2022) to the creative approaches to predictive modeling and data integration put forth by Yang et al. (2021) and Lobato-Delgado et al. (2022). Furthermore, the groundbreaking studies by Haldavnekar et al. (2020) and Zhou et al. (2023) demonstrate the potential of machine learning in

predicting the fate of cancer stem cells and the prognosis of postoperative gastric cancer, respectively, highlighting the ongoing progress in this area and the increasing importance of machine learning-driven approaches in improving cancer patient outcomes.

Methodology

This section discusses the methods employed in this research. All codes were written with python programming language, including Python libraries on Jupyter notebook. Figure 3.1 shows a schematic view of this section.

3.1 Clinical Dataset Collection

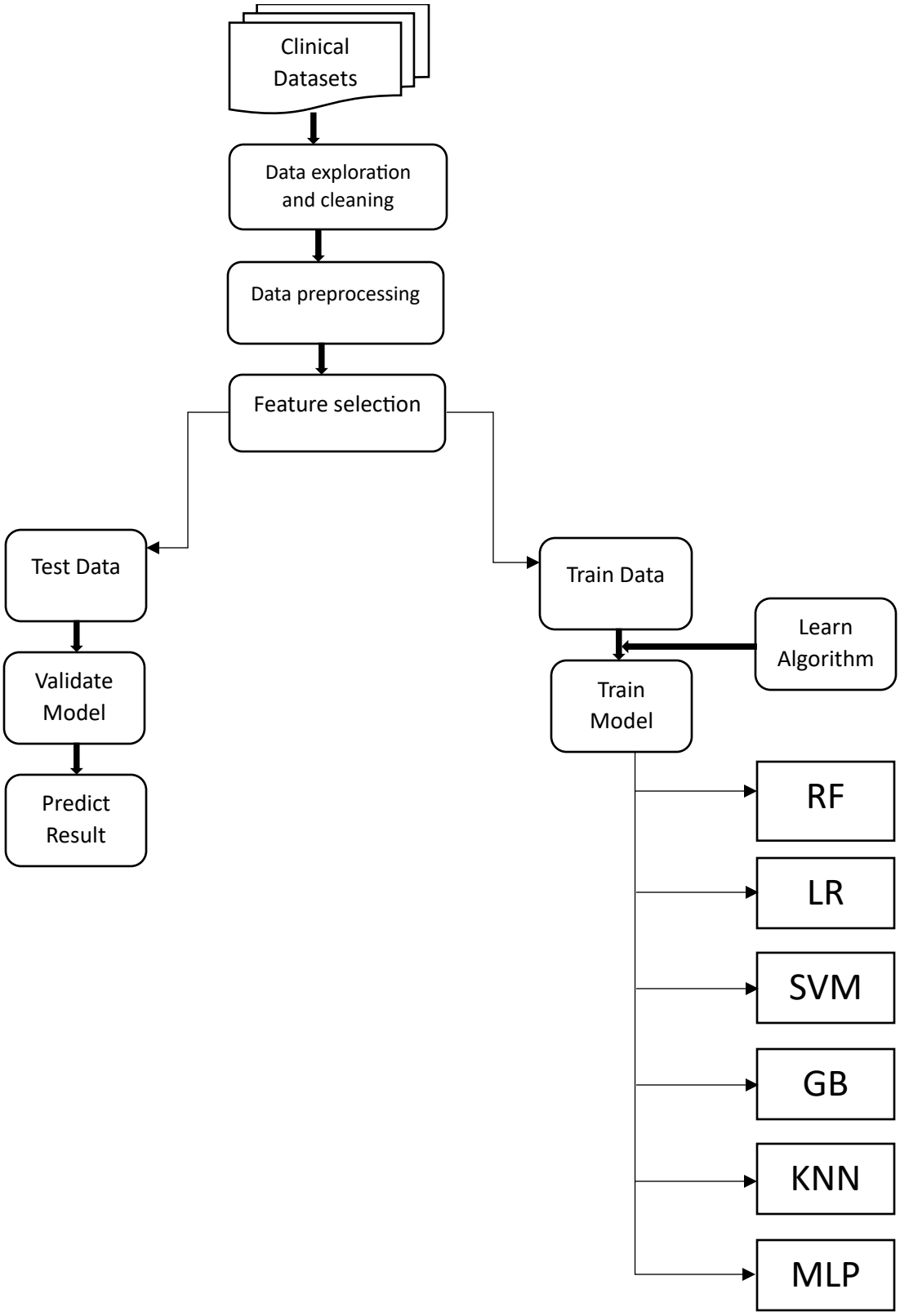
The data collection and preprocessing procedure aimed to obtain miRNA expression data from the Genomic Data Commons (GDC) using the GDC API. The process begins by extracting sample IDs from a manifest file, which serve as key identifiers. A payload is then constructed to query the GDC API, specifying the data type as 'miRNA Expression Quantification' and filtering files based on the obtained sample IDs. A POST request to the GDC API is used to retrieve metadata associated with the selected samples, including important patient information, disease types, and file details.

After obtaining the data, miRNA expression files for the specified samples were retrieved from the GDC API. These downloaded files, organized by sample types, contained the necessary raw genomic information for subsequent analysis. During preprocessing, each miRNA expression file was individually processed. This involved transposing the files to reorient the data and extracting relevant details from the metadata file. A label dictionary (labelDict) was used to assign labels based on associated disease types to improve interpretability.

3.2. Machine learning techniques and optimization methods

Five well-known machine learning methods were considered in this study: KNN, GB, SVM, RF, and LR.

Figure 3. Schematic diagram of the work



3.2.1 Data exploration and cleaning

The dataset's dimensions and characteristics were explored, revealing key insights. The dataset was scrutinized for missing values (A), and the absence of null entries ensured a robust foundation for subsequent analyses. The dataset comprises samples from diverse cancer types. Less-represented groups were eliminated to correct the sample's distributional imbalances (B). Choosing an equal number of records randomly for each class, depending on the minimum class size, was crucial (C).

3.2.2 Data preprocessing

The preprocessing began by separating the feature values (X) and class labels (y) from the original genomic expression dataset, ensuring they were in a format suitable for machine learning in the scikit-learn framework (D). The feature matrix, represented as X, contained gene expression profiles from various samples, while the target vector, denoted as y, indicated the cancer types for each sample. The LabelEncoder was used to translate categorical class labels into numerical values to prepare for efficient model training (E). The dataset was divided into training and test sets after encoding (F). The dataset was divided into test and training sets after encoding. Further analysis of the models during training involved separating the training set into separate subsets for training and validation. By scaling feature values to a range of 0 to 1, data normalization techniques were used to increase the robustness of the model (G). These preprocessing steps ensured the dataset was adequately formatted, encoded, and normalized, laying a strong foundation for developing precise and resilient classification models.

3.2.3 Feature selection

Using the mutual_info_classif function from scikit-learn, we computed mutual information (MI) scores between the normalized feature set and encoded target labels. By quantifying the

informational link between specific target labels and particular features, these MI scores offer valuable insights into the relevance of the data for the classification task. The top 400 features were then chosen (though this number can be changed for experimentation) from a sample of features sorted according to their mutual information scores. Keeping the indices corresponding to the highest-ranking features and sorting the MI scores in descending order were part of the selection procedure (H). The feature sets for the training, validation, and test datasets were then updated using these chosen features. This approach aims to enhance model interpretability and improve classification performance by focusing on the most informative features.

3.2.4 Classification

This research project encompassed the application of six distinct classifiers—RF, LR, SVM, GB, KNN, and NN for cancer classification.

RF, LR, SVM, GB, KNN, and MLP were the six classifiers used in this research project to classify cancer. Each classifier was configured in a framework of one against the other to predict a class from the other. The training dataset was used to train each classifier independently on the chosen feature set. The classifiers included RF, LR (using the' sag' solver with a maximum iteration limit of 5000 to assure convergence), SVM (with a linear kernel), GB, KNN, and MLP. Subsequently, forecasts were generated using the test dataset, and thorough performance assessments were carried out.

3.2.5 Performance metrics

Performance Metrics such as accuracy, precision, recall, and F1-score were computed, providing a holistic assessment of each classifier's discriminative capabilities. The mathematical representation of the measures mentioned above is calculated as follows:

1. Accuracy (ACC) = $\frac{TP+TN}{TP+TN+FP+FN}$ (i)

2. Precision (P) = $\frac{TP}{TP+FP}$ (ii)

3. Recall (R) or Sensitivity or True Positive Rate = $\frac{TP}{TP+FN}$ (iii)

4. F1-Score = $2 \times \frac{P \times R}{P+R}$ (iv)

Where:

TP is the number of true positives (correctly predicted positive samples)

TN is the number of true negatives (correctly predicted negative samples)

FP is the number of false positives (incorrectly predicted positive samples)

FP is the number of false negatives (incorrectly predicted negative samples)

Furthermore, classification reports and confusion matrices were generated, offering insights into the model's performance for each cancer type. The study culminated in a comparative analysis of the classifiers, identifying the most suitable approach for multi-class cancer classification.

3.2.6 Build the neural network model

Using the Keras Sequential API, a NN model was built in this study to categorize various forms of cancer. The model comprised three layers: an output layer with 10 neurons and two hidden layers with 40 and 20 neurons each. Softmax was utilized to translate the model's outputs into probabilities for each type of cancer, boosting its capability to predict many types. The model was trained with a particular learning rate using the Adam optimization method, and its performance was assessed using a particular kind of loss function. A different dataset was used to evaluate the model's learning progress during several training epochs. The model's accuracy in predicting

various cancer kinds was then evaluated using a fresh dataset. During training, the model's performance was shown through graphic representations. Algorithm 1 outlines the key steps performed in this methodology.

Algorithm 1: Cancer Classification using Machine Learning

Data Loading and Exploration:

Input: Dataset file path

Output: DataFrame containing the dataset

Algorithm:

Load the dataset into a DataFrame using the provided file path.

Display the shape of the DataFrame to show the number of rows and columns.

Use the describe method to generate descriptive statistics of the dataset.

Print the names of the first three columns.

Print the name of the last column.

Check for missing values and display the columns with missing values.

Display the distribution of classes in the dataset using a bar chart.

Data Preprocessing:

Input: DataFrame

Output: Preprocessed DataFrame with balanced classes

Algorithm:

Remove columns with missing values.

Visualize the class distribution and remove classes with fewer records.

Balance the dataset by randomly selecting an equal number of records for each class.

Split the dataset into training, validation, and test sets.

Normalize the feature values between 0 and 1 using Min-Max scaling.

Feature Selection:

Input: Training and validation sets

Output: Selected features for training and validation sets

Algorithm:

Compute Mutual Information scores for each feature.

Select the top features based on their scores.

Extract the selected features from the training, validation, and test sets.

Model Training:

Input: Selected features and corresponding labels for training and validation sets

Output: Trained machine learning models

Algorithm:

Train multiple classifiers (RF, LR, SVM, GB, KNN, MLP) using the selected features and labels.

Train a NN model using TensorFlow/Keras with the selected features and labels.

Model Evaluation:

Input: Trained models, test features, and corresponding labels

Output: Performance metrics and visualizations

Algorithm:

Evaluate each classifier's performance using accuracy, precision, recall, and F1-score.

Generate a classification report and confusion matrix for each classifier.

Plot ROC curves for multi-class classification.

Train and evaluate the NN model, plot training and validation accuracy/loss curves.

Results

4.1 Data Exploration and Cleaning

The dataset comprises 2,507 samples and 1,881 features, providing a rich reservoir of genomic information for analysis. The initial examination revealed diverse gene expression levels across samples, indicative of the complex molecular landscape underlying different cancer types. Descriptive statistics offered valuable insights into the distribution and variability of gene expression, setting the stage for further investigation into potential biomarkers and therapeutic targets. Moreover, the absence of missing values underscores the dataset's completeness, facilitating seamless data preprocessing without the need for imputation techniques that could introduce bias.

Analysis revealed the presence of five distinct classes, with liver cancer (920 samples) emerging as the most prevalent, followed by brain and skin cancers with 637 and 455 samples, respectively (Fig 4.1). This distribution provided valuable context for subsequent analyses, highlighting the need to address class imbalances to prevent bias in classification models. Furthermore, identifying and excluding classes with limited representation, such as 'Stomach' and 'unknown' labels with 39 and 22 samples, respectively, demonstrated a proactive approach to data refinement, ensuring the focus remains on the most well-represented and clinically relevant cancer types.

To mitigate class imbalance, we employed a random sampling approach to select an equal number of records for each cancer type based on the minimum count of records per class (Fig 4.2). This balanced sampling strategy enhances the dataset's suitability for machine learning algorithms, reducing the risk of biased predictions and improving model generalization.

Figure 4. Cancer types class distribution

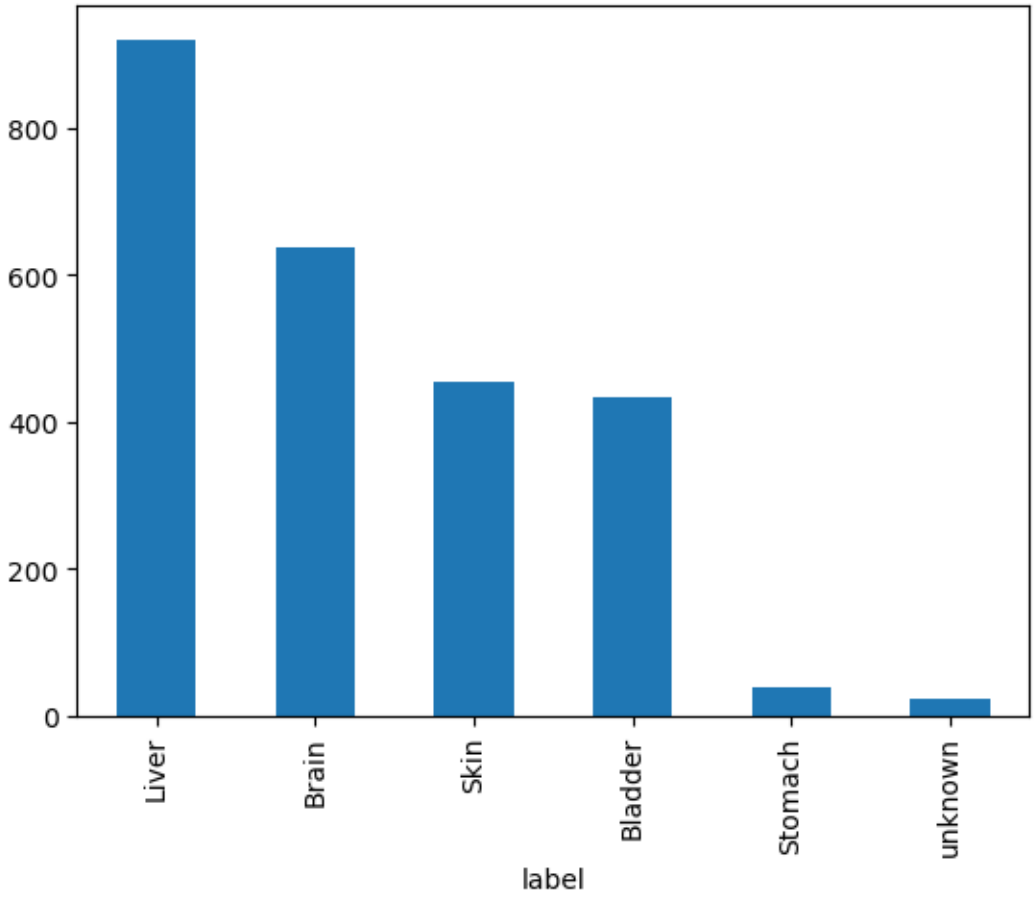
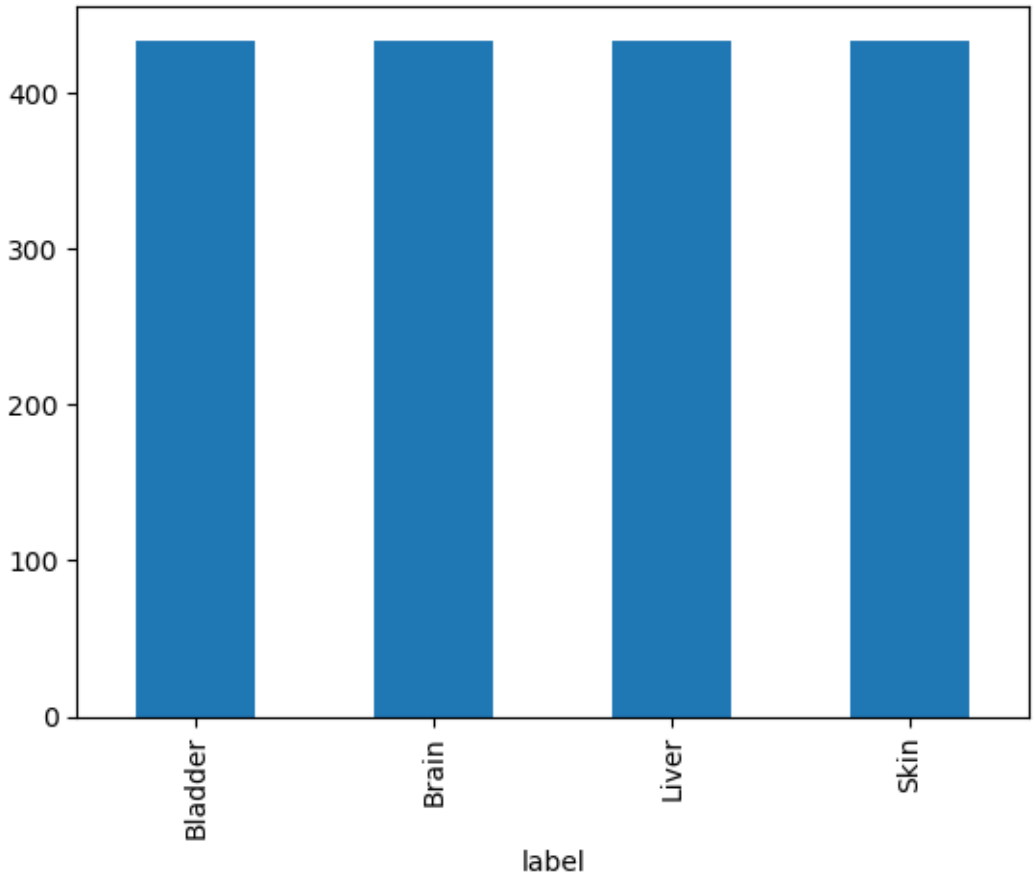


Figure 5. Cancer types class distribution based on random sampling.



4.2 Data Preprocessing

The data preprocessing phase successfully prepared the dataset for machine learning modeling, with meticulous steps ensuring its integrity and compatibility with algorithms. By separating features from the target variable, encoding categorical labels, and stratifying data split into training, validation, and test sets, the groundwork was laid for rigorous model training and evaluation. The features (X) and the target variable (y) were separated to conform to the input requirements of machine learning algorithms. The feature matrix (X) contains 2,446 samples and 1,881 features, while the target array (y) contains corresponding class labels for each sample, encoded numerically for compatibility with machine learning models. Furthermore, data normalization has standardized feature values, mitigating biases and enhancing model performance. These comprehensive preprocessing steps establish a solid foundation for subsequent analysis, facilitating the exploration of intricate gene expression patterns and the development of accurate cancer classification models.

4.3 Feature Selection

Leveraging the mutual information algorithm, the relevance of each feature was computed, enabling the selection of the top N features deemed most pertinent to the classification task. In this case, selecting the top 400 features from the normalized dataset has substantially reduced its dimensionality while preserving the critical information necessary for accurate classification. The resultant feature-selected datasets for training and testing exhibit a refined structure, with dimensions reduced to 1564 samples for training, 490 samples for testing, and 392 samples for validation, each with 400 selected features. These datasets were poised for utilization in machine learning modeling, where the retained features are expected to contribute significantly to the model's predictive performance.

4.4 Performance of Algorithms

4.4.1 Precision

The precision scores obtained from the classification data demonstrated notable variations across different cancer types, reflecting the diverse performance of the machine learning algorithms in correctly identifying instances within each class. Among the classifiers, as seen in Table 4.1a, the RF and NN exhibited exceptional precision scores across all cancer types, with RF achieving precision scores of 0.98 for Bladder, 0.99 for Brain, Liver, and Skin, while NN achieved precision scores of 0.97 for Bladder, 1 for Brain and Liver, and 0.99 for Skin. These results highlighted the robustness of RF and NN in accurately distinguishing between various cancer types, with minimal false positives. Conversely, KNN classifier demonstrated lower precision scores, particularly for Bladder and Skin cancers, achieving scores of 0.94 and 0.96, respectively, indicating a comparatively lower accuracy in correctly identifying instances within these classes.

4.4.2 Recall Scores

Several key observations arise in reviewing the provided classification data, as seen in Table 4.1b, outlining recall scores for different cancer types across various machine learning algorithms. SVM exhibited outstanding performance, with near-perfect recall scores across all cancer types, notably achieving 0.99 for Bladder, Brain, and Skin cancers and an impressive 0.98 for liver cancer detection. RF and GB methods also demonstrated robust performance, consistently achieving high recall scores above 0.98 across all cancer types. However, LR and KNN algorithms exhibited slightly lower recall scores, particularly for bladder and liver cancer detection, with LR achieving 0.96 and 0.98 for bladder and liver cancers, respectively, and KNN scored 0.94 for Bladder and 0.96 for liver

Table 1. Performance Metrics

(a)

	PRECISION			
	Bladder	Brain	Liver	Skin
RF	0.98	0.99	0.99	0.99
LR	0.95	1	0.98	0.99
SVM	0.96	1	1	0.99
GB	0.97	0.99	0.99	1
KNN	0.94	1	0.97	0.96
NN	0.97	1	0.99	0.97

(b)

	RECALL			
	Bladder	Brain	Liver	Skin
RF	0.98	0.99	0.99	0.99
LR	0.96	0.98	0.98	0.99
SVM	0.99	1	0.98	0.99
GB	0.98	0.99	0.99	0.99
KNN	0.94	0.99	0.96	0.99
NN	0.98	0.99	0.98	0.99

(c)

	F1-SCORE			
	Bladder	Brain	Liver	Skin
RF	0.98	0.99	0.99	0.99
LR	0.95	0.99	0.98	0.99
SVM	0.97	1	0.99	0.99
GB	0.97	0.99	0.99	0.99
KNN	0.94	1	0.97	0.97
NN	0.97	1	0.99	0.98

(d)

	OVERALL PERFORMANCE			
	Accuracy	Precision	Recall	F1-score
RF	0.9886	0.9898	0.9898	0.9898
LR	0.9786	0.9797	0.9796	0.9796
SVM	0.9904	0.99	0.9898	0.9898
GB	0.9872	0.9878	0.9878	0.9878
KNN	0.9697	0.9695	0.9694	0.9694
NN	0.9858	0.9859	0.9857	0.9858

cancers.

4.4.3 F1-Score

Table 4.1c provides data on the F1-score metrics. SVM again demonstrated strong performance, achieving high F1-scores consistently across all cancer types, with notable scores of 0.97 for bladder and skin cancers and a perfect score of 1 for brain cancer detection. RF and GB methods also exhibited commendable performance, consistently scoring above 0.97 across all cancer types. LR showed competitive results, particularly for brain and skin cancer detection, with F1-scores of 0.99. However, KNN and NN algorithms, while achieving perfect scores for brain cancer, exhibited slightly lower F1-scores for other cancer types, suggesting room for improvement or optimization in these methods. Overall, these F1-score metrics provided valuable insights into the efficacy of different machine learning algorithms for cancer classification tasks, with SVM, RF, and GB algorithms standing out for their robust performance across various cancer types.

4.4.4 Overall Performance Metrics

Table 4.1d provided an overall performance of the algorithms presenting accuracy, precision, recall, and F1-score metrics for different machine learning algorithms across various cancer types. The SVM stood out with the highest accuracy of 0.9904, as well as precision, recall, and F1-score, all at 0.9898, demonstrating its effectiveness across different evaluation metrics. RF also showed strong performance with an accuracy of 0.9886 and precision, recall, and F1-score, all at 0.9898, making it a competitive choice for cancer classification tasks. LR and GB methods exhibited comparable performance, achieving accuracy scores of 0.9786 and 0.9872, respectively, with precision, recall, and F1-scores around 0.9796 and 0.9878. However, KNN and NN algorithms, while maintaining reasonable accuracy levels, demonstrated slightly lower precision, recall, and F1-scores, indicating potential areas for improvement or optimization in these methods

4.4.5 Confusion Matrix and ROC

The results of the classification model are presented through the confusion matrix, which showcases the predicted labels against the actual labels. Each cell in the matrix represents the number of instances where a predicted label corresponds to an actual label. In this case, the RF classifier performed well, as evidenced by the diagonal elements having high values, indicating correct predictions. RF accurately predicted 94 instances of Bladder, 126 instances of Brain, 176 instances of Liver, and 89 instances of Skin. However, a few misclassifications are evident in off-diagonal elements, such as 1 instance of Brain being misclassified as Bladder and one instance of Skin being misclassified as Liver. To visually represent the confusion matrix, a heatmap was generated using Seaborn, a Python data visualization library (Fig 4.3). The ROC curve for each class was then plotted with a dashed line, labeled to denote the class against the rest, accompanied by its AUC value (Fig 4.4)

The LR classifier, upon analysis, exhibited satisfactory performance, as indicated by predominantly high values along the diagonal elements, signifying accurate predictions as seen in Fig 4.5. The model correctly predicted 92 instances of Bladder, 125 instances of Brain, 174 instances of Liver, and 89 instances of Skin. However, there are a few misclassifications evident in off-diagonal elements, such as three instances of Bladder being misclassified as Liver and one instance

Figure 6. Confusion Matrix using Random Forest Classifier

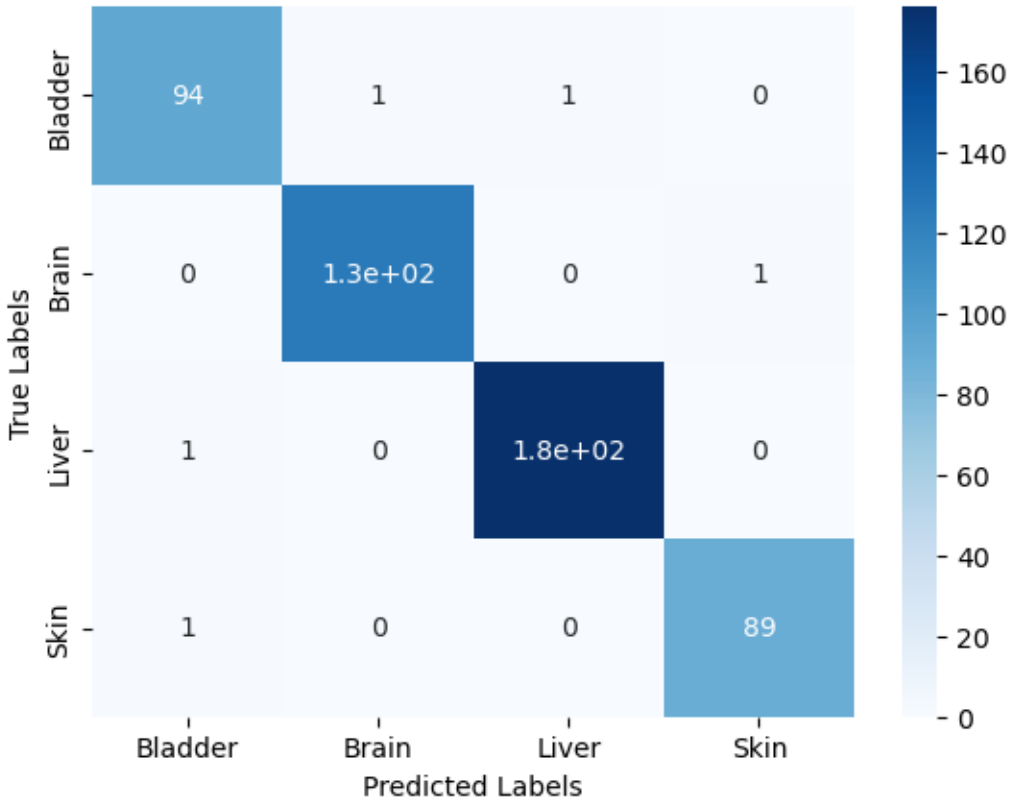


Figure 7. Multi-class ROC Curve using the Random Forest Classifier

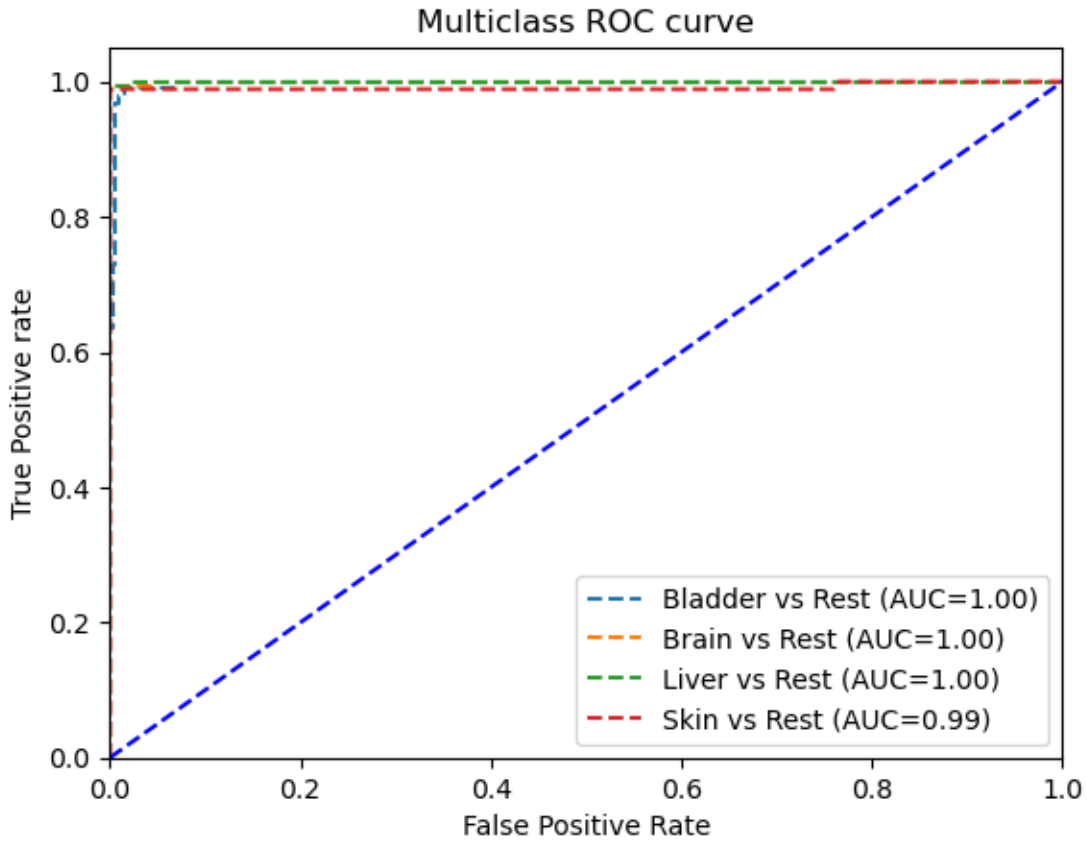
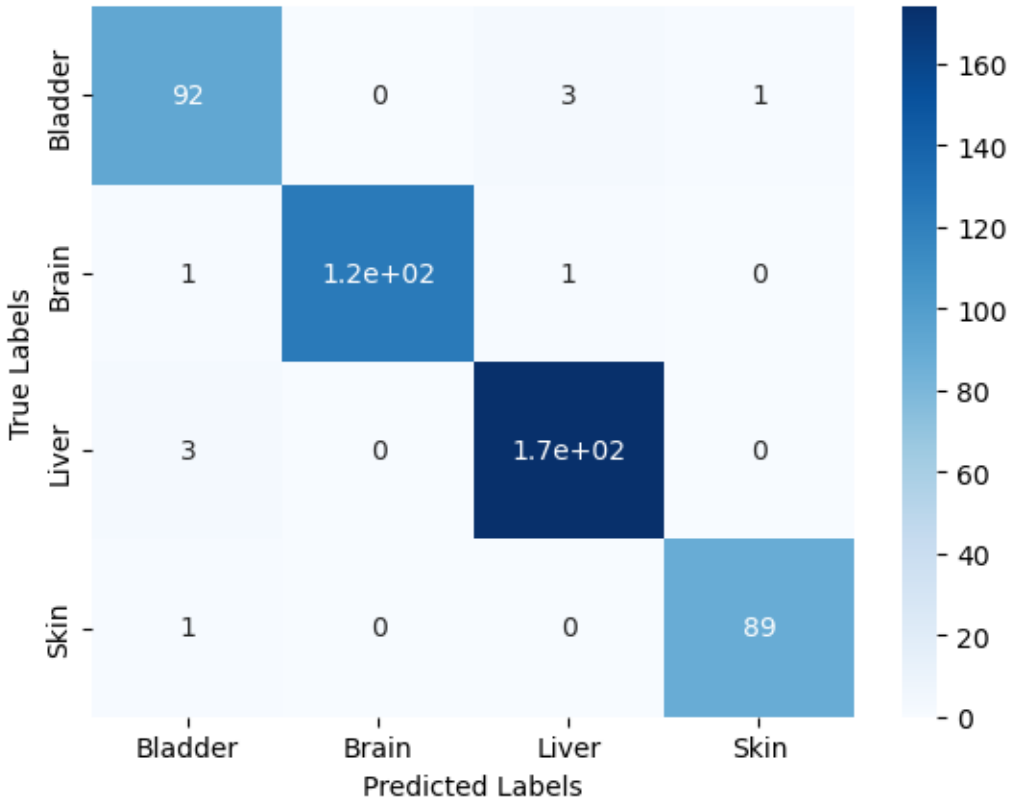


Figure 8.Confusion Matrix using Logistic Regression Classifier



4.5 Neural Network Model Performance

The NN model trained with the specified architecture and parameters achieved impressive results on the dataset. During the training process, the model quickly converged to high accuracy levels on both the training and validation sets, with the training accuracy reaching near-perfect levels and the validation accuracy consistently high. The evaluation of the test set showed that the model's predictions closely matched the actual labels (Fig 4.6), indicating robust performance in unseen data. The loss curves (Fig 4.7) also demonstrated rapid convergence and stabilization, indicating effective learning throughout the training process. This suggests that the model generalizes well to new samples, confirming its effectiveness in classification tasks.

Figure 9. Neural network model performance curve

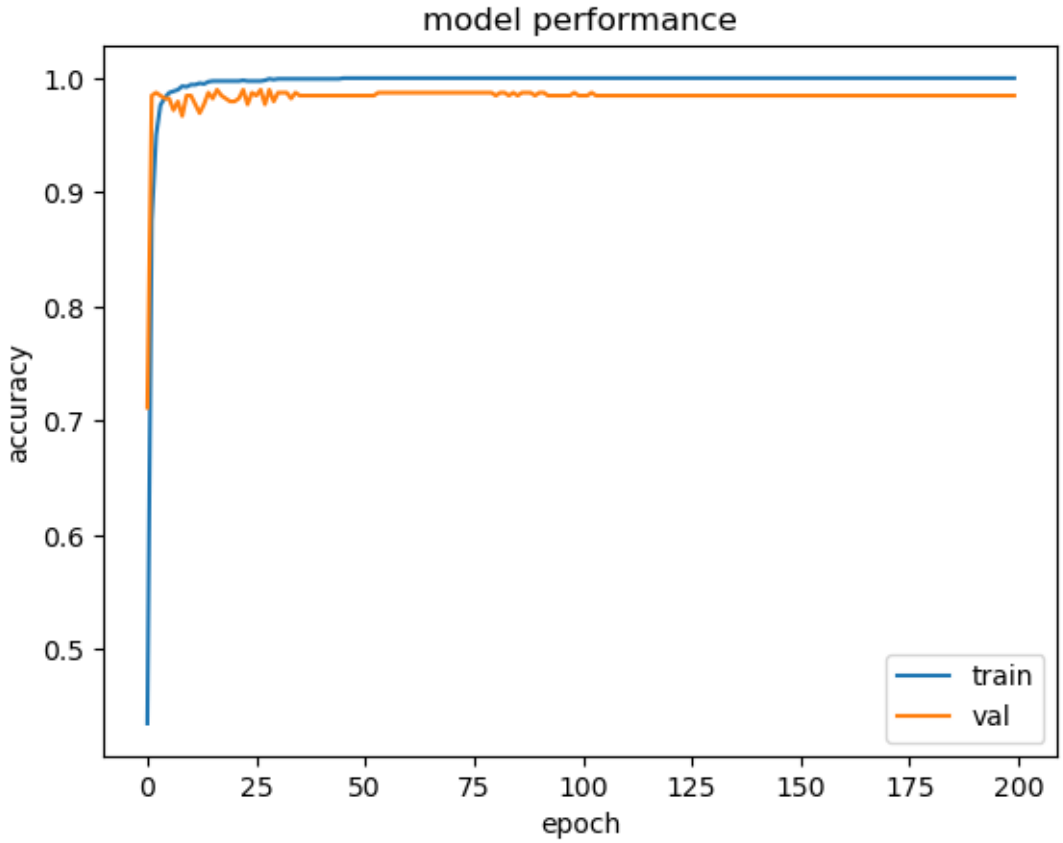
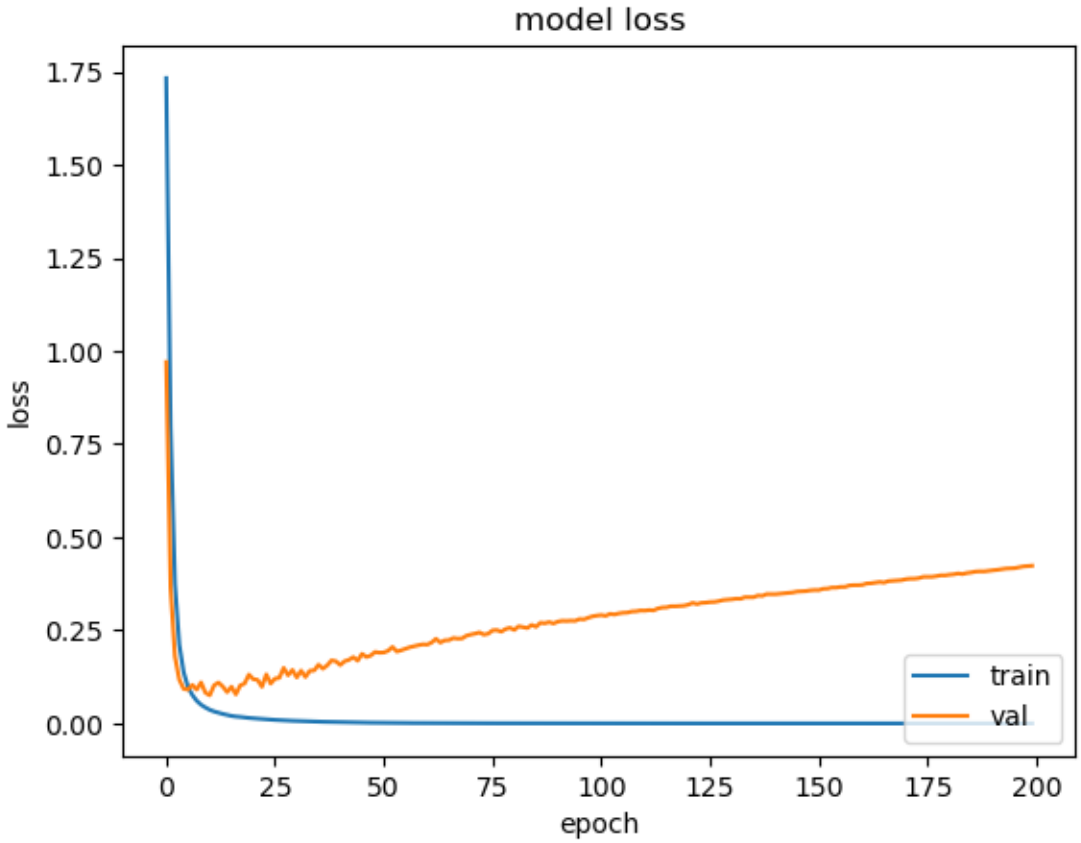


Figure 10. Neural network model loss curve



Discussion

An ever-growing range of multi-class classification problems arises from the investigation of cancer using gene expression research. There is an increasing need for creative and practical solutions to address the shortcomings of traditional machine learning techniques, such as reproducibility and interpretability.

In this study, we have introduced a novel approach for the data-driven classification of multiple cancer categories using RNA-seq. The principal advantage of our approach is in the robustness and ease of use of the top-scoring gene sets, which provide clear limits for interpretation based on biology. Studies like Kaur et al. (2012) and Haibe-Kains et al. (2012) show that recent research has begun to support these simple models.

The study evaluated the efficacy of six machine learning-based classification algorithms for multi-class cancer classification: RF, LR, SVM, GB, KNN and NN. SVM showed the best performance among these algorithms, which is in line with earlier research findings (Wu and Hicks 2021). SVM was found to be superior to other algorithms, such as ANN, DT, NB, and KNN, in a meta-analysis by Nindrea *et al*. (2018). Furthermore, research on categorizing breast cancer using imaging data has also demonstrated the efficacy of SVM (La Forgia *et al.* 2020). Furthermore, SVM's superiority over alternative algorithms for breast cancer classification was established by Wu and Hicks (2021). SVM has demonstrated potential in risk assessment, diagnosis, and outcome prediction in addition to cancer classification (Kourou *et al.* 2015; Heidari *et al.* 2018), indicating its possible clinical utility outside the purview of our study. Although SVM's applications in risk assessment and outcome prediction were not examined in our study, the literature currently in publication (Heidari *et al.* 2018; Kourou *et al.* 2015) demonstrates its
superiority over competing algorithms, which is consistent with our findings.

In a study on liver cancer categorization by Rajesh et al. (2020), RF obtained the most remarkable accuracy of 80.64%. Furthermore, the accuracy of the SVM classifier for a dataset was only 84.6%, according to earlier research. It should be emphasized, nonetheless, that taking into account a more extensive dataset can help to increase this accuracy further. Our results supported this claim, showing that utilizing a 400-feature dataset increased accuracy compared to the study's original dataset, which had only 68 features.

Histopathologists often employ immunohistochemical staining as part of traditional cancer categorization methods, with additional imaging techniques employed (Wu and Hicks 2021; Fanizzi *et al.* 2019). Radiomics is increasingly used as a diagnostic technique (Koçak *et al.* 2019; Santos *et al.* 2019). Although these techniques have demonstrated some efficacy, the machine learning algorithms suggested in this work present a novel method for precisely categorizing cancerous tumors, which may supplement currently available imaging modalities. Furthermore, machine learning algorithms can reduce the likelihood of human error caused by expert tiredness or inexperience, especially when processing medical data more quickly and thoroughly. Early and accurate identification of cancer patients, given the aggressiveness and lethality of the disease, may facilitate timely interventions and ultimately lead to improved outcomes.

Our study demonstrates the potential of machine learning algorithms for classifying various forms of cancer. However, it is critical to recognize the limitations of our research. Even though we analyzed and contrasted the performance of five algorithms, our analysis left out many other classifiers. SVM performs exceptionally well in cancer classification tasks, despite this drawback, according to prior analyses of alternative methods (Wu and Hicks 2021; Heidari *et al.* 2018)

Conclusion

6.1 Conclusions

From the result obtained from this research, the following conclusions were made:

1. The study collected miRNA expression data from GDC, comprising 2,507 samples and 1,881 features. The samples had 920 liver cases, 637 brain cases, 455 skin cases, 435 bladder cases, 39 stomach cases, and 22 unknown cases. The dataset's completeness and integrity were ensured through meticulous preprocessing steps, with no missing values requiring imputation techniques. Descriptive statistics revealed diverse gene expression levels, setting the stage for further exploration into potential biomarkers and therapeutic targets. Class balancing techniques were also implemented to mitigate biases and enhance the dataset's suitability for machine learning algorithms, ultimately improving model generalization.

2. Six machine learning methods, RF, LR, SVM, GB, KNN, and NN, were used in exploring and optimizing dataset characteristics for cancer classification. The dataset's characteristics were fine-tuned through rigorous experimentation to enhance model performance. Stratified data splits were employed to ensure representative training, validation, and testing sets. Moreover, data normalization was applied to standardize feature values, mitigating biases and improving the algorithms' ability to extract meaningful patterns from the data.

3. By computing the relevance of each feature, the top 400 features deemed most pertinent to the classification task were selected. This approach effectively reduced the dataset's dimensionality while preserving critical information necessary for accurate classification. The resulting feature-selected datasets exhibited a refined structure, poised for utilization in machine

learning modeling, where the retained features were expected to contribute significantly to the model's predictive performance.

4. Six trained classifiers were independently evaluated for their performance in cancer classification. Precision, recall, and F1-score metrics were assessed across various cancer types to gauge the algorithms' effectiveness. A notable performance was observed in classifiers such as RF (0.9886, 0.9898, 0.9898, 0.9898) and NN (0.9858, 0.9859, 0.9857, 0.9858) in Accuracy, Precision, Recall, and F1-score respectively, which demonstrated robustness in accurately distinguishing between different cancer types with minimal false positives. SVM also stood out for its outstanding performance, with an accuracy score of 0.9904, reaffirming its efficacy in cancer classification tasks.

5. Lastly, this study concluded that the model exhibited impressive results, quickly converging to high accuracy levels during training. Evaluation of the test set demonstrated the model's ability to generalize well to unseen data, confirming its effectiveness in classification tasks. The rapid convergence and stabilization of loss curves further indicated the model's capacity to capture complex relationships within the data. Overall, the NN model showcased promising potential for improving cancer classification accuracy and contributing to advancements in personalized medicine.

6.2 Contribution to Knowledge

This study significantly contributes to cancer classification and genomic analysis by providing a systematic approach to handling miRNA expression data. The study ensures the integrity and suitability of the dataset for machine learning modeling through meticulous preprocessing techniques, including class balancing strategies and feature selection methodologies. By

addressing class imbalances and selecting informative features based on Mutual Information scores, the study enhances the interpretability of classification models, thereby facilitating a deeper understanding of the underlying biological mechanisms driving cancer subtypes. Furthermore, the comprehensive evaluation of six classifiers offers insights into their comparative performance, guiding researchers in selecting appropriate modeling approaches tailored to specific cancer types and performance metrics.

Moreover, the successful construction and validation of a NN model for cancer classification underscore its efficacy in capturing complex relationships within genomic data. This validation contributes to advancing the application of deep learning techniques in cancer research, promising improved classification accuracy and novel insights into cancer biology.

As chemical engineers embarking on machine learning-based cancer prediction with large-scale clinical data, we harness the power of machine learning algorithms to develop predictive models for early cancer detection and personalized treatment strategies. This effort not only aligns with the core principles of chemical engineering, emphasizing the analysis and manipulation of complex systems, but also underscores the importance of innovation and collaboration in addressing pressing global health challenges. Furthermore, this research venture presents a unique opportunity to apply fundamental chemical engineering concepts such as process optimization and systems analysis to the intricate landscape of medical data analytics. Overall, the study's contributions enhance the methodologies for analyzing genomic data and hold significant implications for precision oncology and personalized medicine initiatives, ultimately advancing our ability to diagnose and treat cancer more effectively.

6.3 Recommendations

Based on the insights gained from this study, several recommendations emerge to steer future research and clinical practice within the realm of cancer classification and genomic analysis;

To have a thorough understanding of cancer biology, researchers ought to investigate the integration of several data types, including protein expression, DNA methylation, mRNA expression, and miRNA expression. Researchers can get new insights into the molecular pathways underlying the spread of cancer and identify prospective treatment targets by utilizing multi-omics data integration techniques.

Furthe research is necessary to improve the interpretability and performance of the model by exploring feature selection techniques other than Mutual Information scores. Methods that can enhance the discriminatory strength of machine learning models and provide supplementary insights into the underlying structure of genomic data include principal component analysis, autoencoder-based dimensionality reduction, and recursive feature elimination.

Furthermore, they must be validated on external datasets to ensure that produced models can be applied to other populations and experimental settings. Research institutions and data consortia working together can make it easier to share data and increase the reproducibility of research findings, which would improve the accuracy and usefulness of machine learning models in healthcare settings.

Future studies should also consider the social and ethical ramifications of incorporating genetic data into treatment. To ensure that genetic research maximizes benefits while reducing risks and downsides, concerns about patient privacy, data security, and informed permission should be carefully considered.

Finally, it is crucial to support the translation of research findings into practice and their sharing with various stakeholders, such as the public, legislators, and healthcare professionals. Researchers can optimize their impact on cancer care and public health by involving stakeholders in the study process and encouraging interdisciplinary cooperation.

References

Abdulqader, Dildar Masood, A Mohsin Abdulazeez, and Diyar Qader Zeebaree. (2020). "Machine learning supervised algorithms of gene selection: A review." *Machine Learning* 62 (03): 233-244.

Abraham, Alison G, Donald D Duncan, Stephen J Gange, and Sheila West. (2009). "Computer-aided assessment of diagnostic images for epidemiological research." *BMC Medical Research Methodology* 9: 1-8.

Akagi, Tomonori, and Masafumi Inomata. (2020). "Essential Updates 2018/2019: Essential advances in surgical and adjuvant therapies for colorectal cancer." *Annals of Gastroenterological Surgery* 4 (1): 39-46.

Alelyani, Salem, Jiliang Tang, and Huan Liu. (2018). "Feature selection for clustering: A review." *Data clustering*: 29-60.

Ali, Md Ramjan, Shah Md Ashiquzzaman Nipu, and Sharfuddin Ahmed Khan. (2023). "A decision support system for classifying supplier selection criteria using machine learning and random forest approach." *Decision Analytics Journal* 7: 100238.

Ali, Najat, Daniel Neagu, and Paul Trundle. (2019). "Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets." *SN Applied Sciences* 1: 1-15.

Allemani, Claudia, Tomohiro Matsuda, Veronica Di Carlo, Rhea Harewood, Melissa Matz, Maja Nikšić, Audrey Bonaventure, Mikhail Valkov, Christopher J Johnson, and Jacques Estève. (2018). "Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries." *The Lancet* 391 (10125): 1023-1075.

Alzubi, Jafar, Anand Nayyar, and Akshi Kumar. 2018. "Machine learning from theory to algorithms: an overview." Journal of physics: conference series.

Awad, Fouad H, Murtadha M Hamad, and Laith Alzubaidi. (2023). "Robust classification and detection of big medical data using advanced parallel K-means clustering, YOLOv4, and logistic regression." *Life* 13 (3): 691.

Awad, Mariette, Rahul Khanna, Mariette Awad, and Rahul Khanna. (2015). "Support vector machines for classification." *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*: 39-66.

Aziz, Norshakirah, Emelia Akashah Patah Akhir, Izzatdin Abdul Aziz, Jafreezal Jaafar, Mohd Hilmi Hasan, and Ahmad Naufal Che Abas. 2020. "A study on gradient boosting algorithms for development of AI monitoring and prediction systems." 2020 International Conference on Computational Intelligence (ICCI).

Bentéjac, Candice, Anna Csörgő, and Gonzalo Martínez-Muñoz. (2021). "A comparative analysis of gradient boosting algorithms." *Artificial Intelligence Review* 54: 1937-1967.

Blandin Knight, Sean, Phil A Crosbie, Haval Balata, Jakub Chudziak, Tracy Hussell, and Caroline Dive. (2017). "Progress and prospects of early detection in lung cancer." *Open biology* 7 (9): 170070.

Burback, Lisa, Suzette Brult-Phillips, Mirjam J Nijdam, Alexander McFarlane, and Eric Vermetten. (2024). "Treatment of posttraumatic stress disorder: a state-of-the-art review." *Current Neuropharmacology* 22 (4): 557-635.

Bychkov, Dmitrii, Nina Linder, Riku Turkki, Stig Nordling, Panu E Kovanen, Clare Verrill, Margarita Walliander, Mikael Lundin, Caj Haglund, and Johan Lundin. (2018). "Deep

learning based tissue analysis predicts outcome in colorectal cancer." *Scientific reports* 8 (1): 3395.

Campbell, Colin, and Yiming Ying. 2022. *Learning with support vector machines*. Springer Nature.

Cengil, Emine, and Ahmet Cinar. 2018. "A deep learning based approach to lung cancer identification." 2018 International conference on artificial intelligence and data processing (IDAP).

CHAUDHURY, SUSHOVAN, SHAMEEK MUKHOPADHYAY, SADEM NABEEL, and DR KBAH. (2021). "A SYSTEMATIC REVIEW OF CAD SYSTEM BASED APPROACH IN DIAGNOSING BREAST CANCER AND ANALYZE EFFECTIVENESS OF MACHINE LEARNING AND DEEP LEARNING ALGORITHMS IN EARLY DETECTION."

Chhikara, Bhupender S, and Keykavous Parang. (2023). "Global Cancer Statistics 2022: the trends projection analysis." *Chemical Biology Letters* 10 (1): 451-451. Figure 1. Adapted with permission from "Chemical Biology Letters," vol. 10, no. 1 (2023): 451.

Crosby, David, Sangeeta Bhatia, Kevin M Brindle, Lisa M Coussens, Caroline Dive, Mark Emberton, Sadik Esener, Rebecca C Fitzgerald, Sanjiv S Gambhir, and Peter Kuhn. (2022). "Early detection of cancer." *Science* 375 (6586): eaay9040.

Dagogo-Jack, Ibiayi, and Alice T Shaw. (2018). "Tumour heterogeneity and resistance to cancer therapies." *Nature reviews Clinical oncology* 15 (2): 81-94.

Dare, Anna J, Gregory C Knapp, Anya Romanoff, Olalekan Olasehinde, Olusola C Famurewa, Akinwumi O Komolafe, Samuel Olatoke, Aba Katung, Olusegun I Alatise, and T Peter Kingham. (2021). "High-burden cancers in Middle-income countries: a review of Prevention and early detection strategies targeting At-risk populations." *Cancer Prevention Research* 14 (12): 1061-1074.

Darweesh, M Saeed, Mostafa Adel, Ahmed Anwar, Omar Farag, Ahmed Kotb, Mohamed Adel, Ayman Tawfik, and Hassan Mostafa. (2021). "Early breast cancer diagnostics based on hierarchical machine learning classification for mammography images." *Cogent Engineering* 8 (1): 1968324.

de Boves Harrington, Peter. (2015). "Support vector machine classification trees." *Analytical chemistry* 87 (21): 11065-11071.

De Mattos-Arruda, Leticia, Regina Mayor, Charlotte KY Ng, Britta Weigelt, Francisco Martínez-Ricarte, Davis Torrejon, Mafalda Oliveira, Alexandra Arias, Carolina Raventos, and Jiabin Tang. (2015). "Cerebrospinal fluid-derived circulating tumour DNA better represents the genomic alterations of brain tumours than plasma." *Nature communications* 6 (1): 8839.

Dhilsath, Fathima M, and S Justin Samuel. (2021). "Hyperparameter tuning of ensemble classifiers using grid search and random search for prediction of heart disease." *Computational Intelligence and Healthcare Informatics*: 139-158.

Dobbelaere, Maarten R, Pieter P Plehiers, Ruben Van de Vijver, Christian V Stevens, and Kevin M Van Geem. (2021). "Machine learning in chemical engineering: strengths, weaknesses, opportunities, and threats." *Engineering* 7 (9): 1201-1211.

Duarte, José Marcio, and Lilian Berton. (2023). "A review of semi-supervised learning for text classification." *Artificial Intelligence Review*: 1-69.

Dunjko, Vedran, and Hans J Briegel. (2018). "Machine learning & artificial intelligence in the quantum domain: a review of recent progress." *Reports on Progress in Physics* 81 (7): 074001.

Erdem, Ebru, and Ferhat Bozkurt. (2021). "A comparison of various supervised machine learning techniques for prostate cancer prediction." *Avrupa Bilim ve Teknoloji Dergisi* (21): 610-620.

Erickson, Bradley J, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L Kline. (2017). "Machine learning for medical imaging." *Radiographics* 37 (2): 505-515.

Fanizzi, Annarita, Teresa Maria Basile, Liliana Losurdo, Roberto Bellotti, Ubaldo Bottigli, Francesco Campobasso, Vittorio Didonna, Alfonso Fausto, Raffaella Massafra, and Alberto Tagliafico. (2019). "Ensemble discrete wavelet transform and gray-level co-occurrence matrix for microcalcification cluster classification in digital mammography." *Applied Sciences* 9 (24): 5388.

Ferlay, Jacques, Murielle Colombet, Isabelle Soerjomataram, Donald M Parkin, Marion Piñeros, Ariana Znaor, and Freddie Bray. (2021). "Cancer statistics for the year 2020: An overview." *International journal of cancer* 149 (4): 778-789.

Ferlay, Jacques, Isabelle Soerjomataram, Morten Ervik, Rajesh Dikshit, Sultan Eser, Colin Mathers, Marise Rebelo, Donald M Parkin, D Forman, and F Bray. (2013). "GLOBOCAN 2012 v1. 0, cancer incidence and mortality worldwide." *Iarc Cancerbase* 11.

Freeman, Elizabeth A, Gretchen G Moisen, John W Coulston, and Barry T Wilson. (2016). "Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance." *Canadian Journal of Forest Research* 46 (3): 323-339.

Gaye, Babacar, Dezheng Zhang, and Aziguli Wulamu. (2021). "Improvement of support vector machine algorithm in big data background." *Mathematical Problems in Engineering* 2021: 1-9.

Ghosh, Sourish, Anasuya Dasgupta, and Aleena Swetapadma. 2019. "A study on support vector machine based linear and non-linear pattern classification." 2019 International Conference on Intelligent Sustainable Systems (ICISS).

Golub, Todd R, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, and Mark A Caligiuri. (1999). "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." *science* 286 (5439): 531-537.

Haier, Joerg, and Juergen Schaefers. (2022). "Economic perspective of cancer care and its consequences for vulnerable groups." *Cancers* 14 (13): 3158.

Heidari, Morteza, Abolfazl Zargari Khuzani, Alan B Hollingsworth, Gopichandh Danala, Seyedehnafiseh Mirniaharikandehei, Yuchen Qiu, Hong Liu, and Bin Zheng. (2018). "Prediction of breast cancer risk using a machine learning approach embedded with a locality preserving projection algorithm." *Physics in Medicine & Biology* 63 (3): 035020.

Huang, Shigao, Jie Yang, Simon Fong, and Qi Zhao. (2020). "Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges." *Cancer letters* 471: 61-71.

Iqbal, Muhammad Javed, Zeeshan Javed, Haleema Sadia, Ijaz A Qureshi, Asma Irshad, Rais Ahmed, Kausar Malik, Shahid Raza, Asif Abbas, and Raffaele Pezzani. (2021). "Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future." *Cancer cell international* 21 (1): 1-11.

Jedy-Agba, Elima, Valerie McCormack, Clement Adebamowo, and Isabel dos-Santos-Silva. (2016). "Stage at diagnosis of breast cancer in sub-Saharan Africa: a systematic review and meta-analysis." *The Lancet Global Health* 4 (12): e923-e935.

Koçak, Burak, Emine Şebnem Durmaz, Ece Ateş, and Özgür Kılıçkesmez. (2019). "Radiomics with artificial intelligence: a practical guide for beginners." *Diagnostic and interventional radiology* 25 (6): 485.

Kong, Xiangpan, Hua Li, and Zhengxue Han. (2019). "The diagnostic role of ultrasonography, computed tomography, magnetic resonance imaging, positron emission tomography/computed tomography, and real-time elastography in the differentiation of benign and malignant salivary gland tumors: a meta-analysis." *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology* 128 (4): 431-443. e1.

Kourou, Konstantina, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. (2015). "Machine learning applications in cancer prognosis and prediction." *Computational and structural biotechnology journal* 13: 8-17.

La Forgia, Daniele, Annarita Fanizzi, Francesco Campobasso, Roberto Bellotti, Vittorio Didonna, Vito Lorusso, Marco Moschetta, Raffaella Massafra, Pasquale Tamborra, and Sabina Tangaro. (2020). "Radiomic analysis in contrast-enhanced spectral mammography for predicting breast cancer histological outcome." *Diagnostics* 10 (9): 708.

Lee, Howard, and Yi-Ping Phoebe Chen. (2015). "Image based computer aided diagnosis system for cancer detection." *Expert Systems with Applications* 42 (12): 5356-5365.

Lee, Yeuan Ting, Yi Jer Tan, and Chern Ein Oon. (2018). "Molecular targeted therapy: Treating cancer with specificity." *European journal of pharmacology* 834: 188-196.

Li, Jian. (2022). "Digestive cancer incidence and mortality among young adults worldwide in 2020: A population-based study." *World Journal of Gastrointestinal Oncology* 14 (1): 278.

Liotta, Lance A. (1992). "Cancer cell invasion and metastasis." *Scientific American* 266 (2): 54-63.

Llewellyn, Henry, Jane Neerkin, Lewis Thorne, Elena Wilson, Louise Jones, Elizabeth L Sampson, Emma Townsley, and Joseph TS Low. (2018). "Social and structural conditions for the avoidance of advance care planning in neuro-oncology: a qualitative study." *BMJ open* 8 (1).

Malik, Alok, and Bradford Tuckfield. 2019. *Applied unsupervised learning with R: Uncover hidden relationships and patterns with k-means clustering, hierarchical clustering, and PCA*. Packt Publishing Ltd.

Mankoff, David A. (2007). "A definition of molecular imaging." *The Journal of Nuclear Medicine* 48 (6): 18N.

Mariotto, Angela, Jinani Jayasekerea, Valentina Petkov, Clyde B Schechter, Lindsey Enewold, Kathy J Helzlsouer, Eric J Feuer, and Jeanne S Mandelblatt. (2020). "Expected monetary impact of oncotype DX score-concordant systemic breast cancer therapy based on the TAILORx trial." *JNCI: Journal of the National Cancer Institute* 112 (2): 154-160.

Merlin, D, and JGR Sathiaseelan. (2021). "Improved Classification Accuracy for Identification of Cervical Cancer."

Miller, Kimberly D, Miranda Fidler-Benaoudia, Theresa H Keegan, Heather S Hipp, Ahmedin Jemal, and Rebecca L Siegel. (2020). "Cancer statistics for adolescents and young adults, 2020." *CA: a cancer journal for clinicians* 70 (6): 443-459.

Murali, Nikitha, Ahmet Kucukkaya, Alexandra Petukhova, John Onofrey, C Bishop, V Patel, E Shortliffe, M Stefanelli, A Esteva, and B Kuprel. (2020). "Supervised machine learning in oncology: a clinician's guide." *Digestive disease interventions* 4 (01): 073-081.

Naeem, Samreen, Aqib Ali, Sania Anam, and Muhammad Munawar Ahmed. (2023). "An Unsupervised Machine Learning Algorithms: Comprehensive Review." *Int. J. Comput. Digit. Syst*.

Nenclares, P, and KJ Harrington. (2020). "The biology of cancer." *Medicine* 48 (2): 67-72.

Niknejad, Ali, and Dobrila Petrovic. (2013). "Introduction to computational intelligence techniques and areas of their applications in medicine." *Med Appl Artif Intell* 51: 201.

Nindrea, Ricvan Dana, Teguh Aryandono, Lutfan Lazuardi, and Iwan Dwiprahasto. (2018). "Diagnostic accuracy of different machine learning algorithms for breast cancer risk calculation: a meta-analysis." *Asian Pacific journal of cancer prevention: APJCP* 19 (7): 1747.

Organization, World Health. (2018). "WHO guidelines for the pharmacological and radiotherapeutic management of cancer pain in adults and adolescents."

Pradhan, Kanchan Sitaram, Priyanka Chawla, and Rajeev Tiwari. (2023). "HRDEL: High ranking deep ensemble learning-based lung cancer diagnosis model." *Expert Systems with Applications* 213: 118956.

Priyanka, and Dharmender Kumar. (2020). "Decision tree classifier: a detailed survey." *International Journal of Information and Decision Sciences* 12 (3): 246-269.

Quatrini, Elena, Francesco Costantino, Giulio Di Gravio, and Riccardo Patriarca. (2020). "Machine learning for anomaly detection and process phase classification to improve safety and maintenance activities." *Journal of Manufacturing Systems* 56: 117-132.

Raffle, Angela E, and JA Muir Gray. 2019. *Screening: evidence and practice*. Oxford University Press, USA.

Rajesh, Sanapala, Nurul Amin Choudhury, and Soumen Moulik. 2020. "Hepatocellular carcinoma (HCC) liver cancer prediction using machine learning algorithms." 2020 IEEE 17th India Council International Conference (INDICON).

Rezaeipanah, Amin, and Gholamreza Ahmadi. (2022). "Breast cancer diagnosis using multi-stage weight adjustment in the MLP neural network." *The Computer Journal* 65 (4): 788-804.

Roychowdhury, Sameek, Matthew K Iyer, Dan R Robinson, Robert J Lonigro, Yi-Mi Wu, Xuhong Cao, Shanker Kalyana-Sundaram, Lee Sam, O Alejandro Balbin, and Michael J Quist. (2011). "Personalized oncology through integrative high-throughput sequencing: a pilot study." *Science translational medicine* 3 (111): 111ra121-111ra121.

Saba, Tanzila. (2020). "Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges." *Journal of Infection and Public Health* 13 (9): 1274-1289.

Santos, Marcel Koenigkam, José Raniery Ferreira Júnior, Danilo Tadao Wada, Ariane Priscilla Magalhães Tenório, Marcello Henrique Nogueira-Barbosa, and Paulo Mazzoncini de Azevedo Marques. (2019). "Artificial intelligence, machine learning, computer-aided diagnosis, and radiomics: advances in imaging towards to precision medicine." *Radiologia brasileira* 52: 387-396.

Sarker, Iqbal H. (2021). "Machine learning: Algorithms, real-world applications and research directions." *SN computer science* 2 (3): 160.

Sawicki, Tomasz, Monika Ruszkowska, Anna Danielewicz, Ewa Niedźwiedzka, Tomasz Arłukowicz, and Katarzyna E Przybyłowicz. (2021). "A review of colorectal cancer in terms of epidemiology, risk factors, development, symptoms and diagnosis." *Cancers* 13 (9): 2025.

Schiffman, Joshua D, Paul G Fisher, and Peter Gibbs. (2015). "Early detection of cancer: past, present, and future." *American Society of Clinical Oncology Educational Book* 35 (1): 57-65.

Seyfried, Thomas N, and Leanne C Huysentruyt. (2013). "On the origin of cancer metastasis." *Critical Reviews™ in Oncogenesis* 18 (1-2).

Shinji, Seiichi, Takeshi Yamada, Akihisa Matsuda, Hiromichi Sonoda, Ryo Ohta, Takuma Iwai, Koki Takeda, Kazuhide Yonaga, Yuka Masuda, and Hiroshi Yoshida. (2022). "Recent advances in the treatment of colorectal cancer: A review." *Journal of Nippon Medical School* 89 (3): 246-254.

Shokrzade, Amin, Mohsen Ramezani, Fardin Akhlaghian Tab, and Mahmud Abdulla Mohammad. (2021). "A novel extreme learning machine based kNN classification method for dealing with big data." *Expert Systems with Applications* 183: 115293.

Singh, Sanjay Kumar, Amit Sinha, Harikesh Singh, Aniket Mahanti, Abhishek Patel, Shubham Mahajan, Amit Kant Pandit, and Vijayakumar Varadarajan. (2023). "A novel deep learning-based technique for detecting prostate cancer in MRI images." *Multimedia Tools and Applications*: 1-15.

Soda, Narshone. (2021). "Advanced Liquid Biopsy Technologies for Circulating Cancer Biomarker Detection."

Song, Yan-Yan, and LU Ying. (2015). "Decision tree methods: applications for classification and prediction." *Shanghai archives of psychiatry* 27 (2): 130.

Strickland, Jeffrey. 2017. *Logistic regression inside and out*. Lulu. com.

Thandra, Krishna Chaitanya, Adam Barsouk, Kalyan Saginala, John Sukumar Aluru, and Alexander Barsouk. (2021). "Epidemiology of lung cancer." *Contemporary Oncology/Współczesna Onkologia* 25 (1): 45-52.

Topol, Eric J. (2019). "High-performance medicine: the convergence of human and artificial intelligence." *Nature medicine* 25 (1): 44-56.

Tran, Linda, Jin-Fen Xiao, Neeraj Agarwal, Jason E Duex, and Dan Theodorescu. (2021). "Advances in bladder cancer biology and therapy." *Nature Reviews Cancer* 21 (2): 104-121.

UK, Cancer Research. 2020. "Early Detection." https://www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-important.

Valery, Patricia C, Mathieu Laversanne, Paul J Clark, Jessica L Petrick, Katherine A McGlynn, and Freddie Bray. (2018). "Projections of primary liver cancer to 2030 in 30 countries worldwide." *Hepatology* 67 (2): 600-611.

Ward, Elizabeth M, Recinda L Sherman, S Jane Henley, Ahmedin Jemal, David A Siegel, Eric J Feuer, Albert U Firth, Betsy A Kohler, Susan Scott, and Jiemin Ma. (2019). "Annual report to the nation on the status of cancer, featuring cancer in men and women age 20–49 years." *JNCI: Journal of the National Cancer Institute* 111 (12): 1279-1297.

Wu, Jiande, and Chindo Hicks. (2021). "Breast cancer type classification using machine learning." *Journal of personalized medicine* 11 (2): 61.

Yang, Xiangli, Zixing Song, Irwin King, and Zenglin Xu. (2022). "A survey on deep semi-supervised learning." *IEEE Transactions on Knowledge and Data Engineering*.

Zhang, Yanru, and Ali Haghani. (2015). "A gradient boosting method to improve travel time prediction." *Transportation Research Part C: Emerging Technologies* 58: 308-324.

Zhang, Zhenwei, and Ervin Sejdić. (2019). "Radiological images and machine learning: trends, perspectives, and prospects." *Computers in biology and medicine* 108: 354-370.