Development of an Interpretable Ensemble Model to Predict Immune Checkpoint Inhibitor Therapy Response in Bladder Cancer Patients

by

Syed Ayman Shahbad
22101764
Sadia Ruhama
22141016
Ahnaf Warid
24241270

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

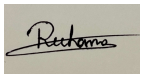
Department of Computer Science and Engineering
Brac University
January 2025

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:



Sadia Ruhama
22141016

Syed Ayman Shahbad
22101764

Ahnaf Warid
24241270

Approval

The thesis/project titled "Development of an Interpretable Ensemble Model to Predict Immune Checkpoint Inhibitor Therapy Response in Bladder Cancer Patients" submitted by

1. Syed Ayman Shahbad (22101764)
2. Sadia Ruhama (22141016)
3. Ahnaf Warid (24241270)

Of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 23, 2025.

Examining Committee:

Supervisor:
(Member)

Dr. Md Ashraful Alam
Associate Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD
Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Sadia Hamid Kazi
Chairperson and Associate Professor
Department of Computer Science and Engineering
BRAC University

Abstract

Identifying robust biomarkers for bladder cancer and immunotherapy continues to be a challenge in the field of computational biology, as pertinent datasets often suffer from the "curse of dimensionality." In this study, we explore the effectiveness of ensemble feature selection methods to discover potential and validate existing biomarkers that are predictive of sensitivity to immune checkpoint inhibitor (ICI) therapy, administered using Pembrolizumab. The thesis utilizes the GSE111636 dataset with 11 samples and over 70,000 features to simulate the high-dimensionality problem and uses a tailored ensemble modeling approach to tackle it. To be more specific, the research applies four popular feature selection algorithms - Support Vector Machine Recursive Feature Elimination (SVM-RFE), Random Forest Recursive Feature Elimination (RF-RFE), Mutual Information (MI), and Logistic Regression Recursive Feature Elimination (LR-RFE) - both individually and in ensembles. To aggregate the results of the ensemble models, a frequency-based majority voting system is adopted. Furthermore, keeping in mind the limitations of the dataset, Leave-One-Out Cross-Validation (LOOCV) is also used to improve on the generalizability and robustness and reduce bias of the feature selection methods. Two deep learning models - one based on neural networks with L1 Regularization and another Concrete Autoencoder - were also used to carry out feature selection. A final comparison between all the models demonstrated that ensemble models outperformed both single models and deep learning models by selecting a greater number of biologically relevant gene features, most of which have been previously implicated in bladder cancer progression and immune response, while also showcasing more stable feature selection metrics. Thus, this study highlights the utility of ensemble modeling in low sample count, high-dimensional gene expression datasets, underscoring its potential for advancing biomarker research in oncology.

Keywords: Bladder Cancer; Ensemble Feature Selection; Biomarkers; High-Dimensional Data; Immune Checkpoint Inhibitor (ICI) Therapy

Acknowledgement

First and foremost, all praise to the Almighty Allah, whose generosity allowed us to complete our thesis within the given time frame.

We would also like to express our heartfelt gratitude to Dr. Md. Ashraful Alam of the School of Computer Science and Engineering at BRAC University; his contribution in the synthesis of this thesis was paramount. Dr. Alam was always approachable and welcoming whenever we encountered any difficulties, steering us in the right direction while consistently allowing this thesis to be our own work.

Lastly, we want to convey gratitude towards the faculty members of the Department of Computer Science and Engineering, whose guidance throughout our undergraduate studies helped us conduct this research.

Sadia Ruhama and Syed Ayman Shahbad contributed equally to this thesis.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

BC Bladder Cancer

$DNNs$ Deep Neural Networks

$FFPE$ Formalin-Fixed Paraffin-Embedded

GEO Gene Expression Omnibus

IC Immune Checkpoint Inhibitor

$LOOCV$ Leave-One-Out Cross-Validation

LR Logistic Regression

$LR-RFE$ Logistic Regression Recursive Feature Elimination

MI Mutual Information

RF Random Forest

$RF-RFE$ Random Forest Recursive Feature Elimination

SVM Support Vector Machine

$SVM-RFE$ Support Vector Machine Recursive Feature Elimination

Chapter 1

Introduction

1.1 Background

Bladder cancer (BC) is a commonly diagnosed malignancy and poses a significant health concern due to its prevalence, recurrence, and treatment challenges. Some of the most common types of BC are urothelial carcinoma (otherwise known as transitional cell carcinoma), squamous cell carcinoma, and adenocarcinoma [45]. Urothelial carcinomas are by far the most widely diagnosed, responsible for over 90 percent of all bladder cancer cases [40]. The World Health Organization's staging of bladder tumors includes low malignancy potential carcinoma in situ or TA, which accounts for 70 percent of all diagnoses. The next stage is the low-grade stage, T1, and T2a, where the tumor has developed into a deeper layer and penetrated the lamina propria. The final stage is the high-grade stage T2b, T3, or T4, which is even deeper and is caused when the tumor goes through the local or lymphatic spread, greatly complicating prognosis and treatment processes [2].
The diagnostic process of BC begins with obtaining a biopsy of the tumor within the bladder, after which the tissue is examined for malignancy. If there is a positive response, the tumor is staged with a CT scan and observed more closely to see whether it has spread and which layers of the urinary bladder are involved [52]. The treatment method is then chosen accordingly; if the tumor detected is in the high-grade stage or other forms of treatment, such as chemotherapy, radiation therapy, etc., do not yield favorable results, then the patient is administered stronger forms of treatment such as immunotherapy. Immunotherapy itself is of various forms, of which Immune Checkpoint Inhibitor therapy is most commonly used due to its favorable response [73]. However, ICI therapy poses greater financial pressure than other cancer treatments, with Pembrolizumab costing up to US$104,244 per year of treatment [17].

1.2 Rational of the Study or Motivation

In the context of advanced-stage BC treatment, the emergence of ICIs such as pembrolizumab has renewed therapeutic hope. However, although ICIs have shown remarkable effectiveness, its effects have only been observed in a subset of patients. Matters are further complicated given the high cost of ICI treatment, making the identification of reliable biomarkers imperative in order to guide patients towards quick and appropriate treatment choices [16].

Machine Learning (ML) models have been extensively used for cancer diagnosis, prognosis, and treatment outcome prediction, including that of ICI. These models generate a mapped output from the input data, and this is conducted via a four-step process that involves (i) data collection, (ii) feature design, (iii) model training, and (iv) model testing [57]. Through this process, these models learn from datasets, which may range from medical images to gene data and clinical records; the use of these models enhances cancer diagnosis and the prediction of treatment responses. More recently, deep neural networks (DNNs), a subset of ML, have been used in various aspects of BC treatment and management, achieving spectacular results but with some caveats. For instance, DNNs learn as a network of interconnected nodes or 2 "neurons", resulting in the requirement of enormous datasets, obtaining which is both computationally and monetarily expensive, especially in the field of biomedicine [57]. Biomedical datasets often suffer from the "curse of dimensionality" issue, complicating model generalization and increasing the risk of overfitting when using single ML models and more so when using DNNs. Furthermore, the effectiveness of immune checkpoint inhibitors is intricately influenced by micro-factors that impact the overall treatment outcome [52]. Thus, the results of models used in treatment outcome prediction most often require interpretation, which is hindered by the "black box" approach of DNNs [57].
On the other hand, ensemble models offer a practical and robust alternative, leveraging multiple algorithms to improve feature selection stability and reduce model bias even when working with high-dimensional datasets [38]. Thus, our study aims to harness the capabilities of ensemble-based feature selection in order to outperform individual or deep learning models in small-sample, high-dimensional contexts. By applying this approach to gene expression data from pembrolizumab-treated BC patients, we aim to find new and validate existing biologically relevant and meaningful biomarkers for BC and ICI therapy sensitivity, while also showcasing the broader applicability of ensemble modeling in cancer research.

1.3 Problem Statement

ICI therapy, administered through drugs such as Pembrolizumab, has emerged as a promising treatment for later-stage bladder cancer. However, its effectiveness is limited to a small subset of patients. Coupled with the high cost and potential side effects of ICI therapy, the need for reliable biomarkers that can predict therapeutic response is crucial. Furthermore, as clinical datasets are expensive and time-consuming to create, there is an urgent need for more advanced models that can identify potential biomarkers even when working with a minimal dataset. This high dimensionality and low sample size characteristic of biomedical datasets often causes existing models, including deep learning models, to struggle to generate results that are clinically applicable; while simple single model approaches suffer from algorithmic bias and a lack of generalizability, DNNs suffer from overfitting and a lack of interpretability due to their "black box" nature. This study addresses this gap by employing an ensemble-based machine learning framework that uses the "wisdom of crowds" ideology to combine multiple feature selection methods through frequency-based majority voting to improve the reliability of the final feature set, finally validating the feature set using the biological relevance of predictive gene signatures.

1.4 Objective

The primary objective of this research is to develop an ensemble model that can identify probable biomarkers predictive of response to Pembrolizumab ICI therapy in BC patients and concurrently validate the reliability of these identified biomarkers. By relying on an ensemble-based approach that integrates multiple feature selection methods through frequency-based majority voting, this research aims to overcome the limitations of high-dimensional, low-sample-size datasets, highlighting the role of ensemble modeling in future biomarker research, especially those carried out on limited data.

1.5 Methodology in Brief

The dataset used in this study was obtained from the Gene Expression Omnibus (GEO) database repository with the accession number GSE111636, titled "Identification of potential biomarkers of response and resistance to programmed cell death-1 blockade in patients with advanced urothelial tumors." This dataset comprises gene expression data of bladder cancer patients undergoing immunotherapy with pembrolizumab (MK-3475), which is an anti-PD-1 immune checkpoint inhibitor (ICI). A total of 11 RNA samples were profiled using the Affymetrix GeneChip Human Transcriptome Array 2.0 (HTA2.0) platform (GPL17586), allowing for high-resolution detection of over 70,523 transcript clusters. The samples were classified based on their clinical response to therapy as either responders (responding positively to ICI therapy) or progressors (not showing a positive response to ICI therapy); out of the 11 samples, 6 were labeled as responders and 5 as progressors.

Preprocessing the dataset began with annotating the raw expression matrix provided in the GSE111636 dataset, which involved mapping the probe identifiers (Affymetrix HTA 2.0 transcript cluster IDs) with their corresponding gene names using the official annotation file available at the Affymetrix website. This annotation enhanced the interpretability of the dataset as a whole. Furthermore, to aid our machine learning pipeline and model input format, we transposed the matrix and added a new column to serve as the target label; samples corresponding to patients who responded to pembrolizumab therapy were labeled as 1 (responders), and non-responders (progressors) were labeled as 0. Finally, after dropping duplicates, rows with null values, and carrying out percentile-based variance thresholding, the dataset shape changed from (11, 70,523) to (11, 1513).

To identify the genes associated with patient response to pembrolizumab, we used four feature selection models based on popular algorithms - Support Vector Machine Recursive Feature Elimination (SVM-RFE), Random Forest Recursive Feature Elimination (RF-RFE), Mutual Information (MI), and Logistic Regression Recursive Feature Elimination (LR-RFE) - in both single and ensemble feature selection. For all models, we utilized Leave-One-Out Cross-Validation (LOOCV) and Stratified K-fold CV to ensure robustness and generalizability, given the limited sample size. In the case of single-model approaches, we extracted the feature list from the final LOOCV fold, selecting approximately 30 top-ranked genes for each method. For ensemble models, we employed an unweighted frequency-based majority voting scheme across all LOOCV iterations, retaining approximately 30 genes that consistently appeared across folds. After feature selection, another layer of weighted

frequency-based majority voting was used to create a final feature set. Finally, the biological relevance of the gene set obtained was compared and validated. Figure 1.1 highlights the outline of this paper.

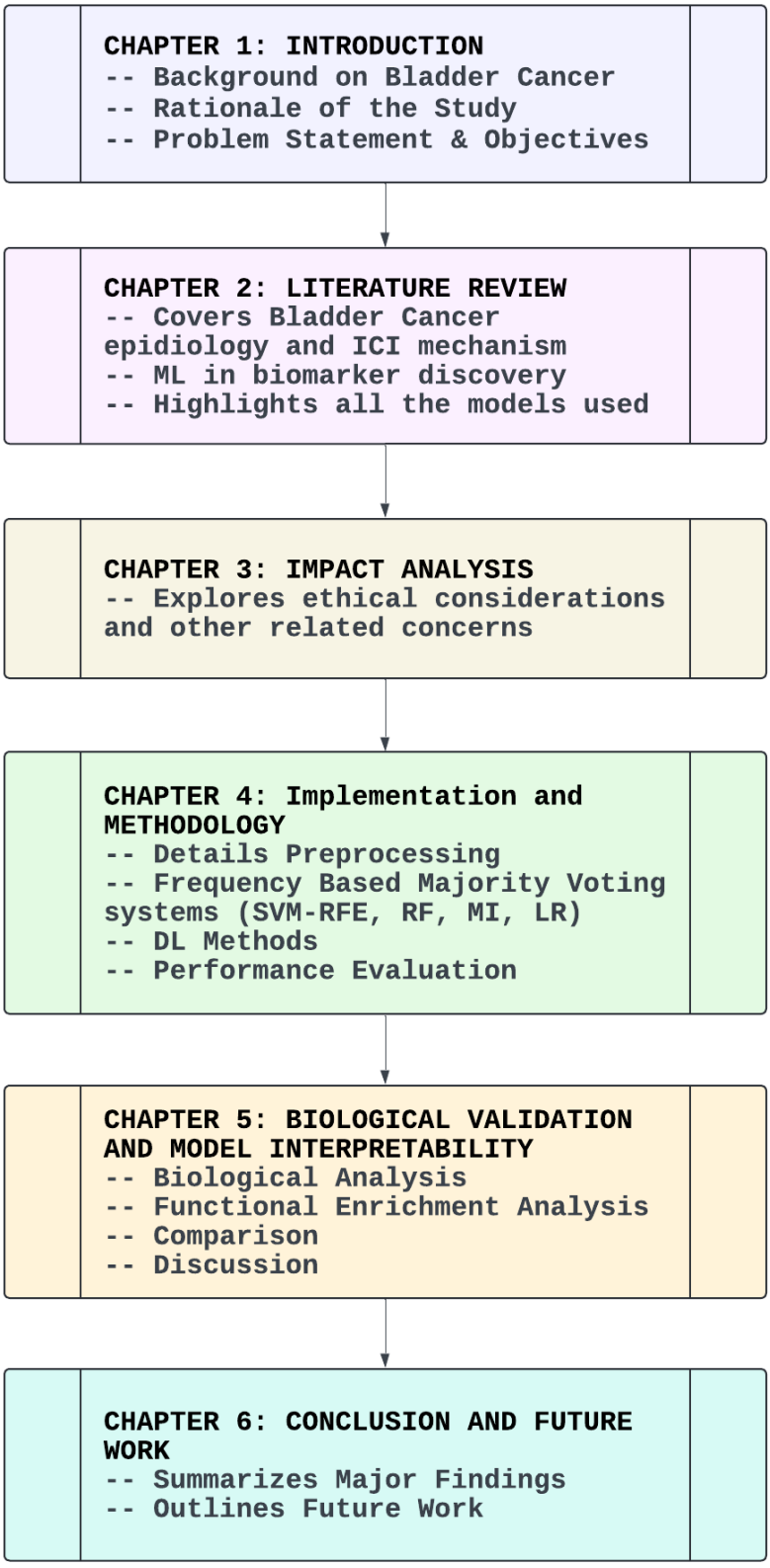


Figure 1.1: Outline

1.6 Scopes and Challenges

Our study was conducted using the GSE111636 dataset from NCBI's Gene Expression Omnibus (GEO) database, which comprises formalin-fixed, paraffin-embedded (FFPE) tissue samples of 11 patients diagnosed with BC. One of the important aspects of this dataset is that all the patients were treated with the same drug, Pembrolizumab, also known as MK-3475. This drug blocks the PD-1 protein on T-cells, which in turn makes the PD-L1 protein produced by the cancer cells unable to bind to PD-1 [18]. As a result, T cells can then actively attack the cancerous cells. Since this dataset is specific to a particular patient cohort, it enables a better understanding of immunotherapy biomarkers that are linked to how patients respond or resist to ICI treatment.

However, due to the nature of our problem and dataset, our research faced certain drawbacks. As there are only 11 patient samples and 70,523 features in our dataset, the statistical metrics of the models applied have increased variance. This is exacerbated by the use of LOOCV, which is known to reduce bias but also introduce higher variance in performance indicators. To acknowledge this drawback, the biological relevance of the selected genes has been strongly analyzed.

Chapter 2

Literature Review and Requirement Analysis

2.1 Bladder Cancer: Epidemiology, Risk Factors, and Treatment Challenges

Bladder cancer is the 9th most commonly occurring cancer in the world, with global trends and patterns suggesting tobacco smoking to be the primary risk factor [20]. Furthermore, age and sex are also determining factors of the predisposition towards getting bladder cancer, as 90 percent of bladder cancers are diagnosed after the age of 55, and men are four times as likely to get bladder cancer during their lifetime as women [39]. In addition, bladder cancer is the 13th most likely cancer to lead to death, with the average survival rate of patients over the course of 5 years being 77 percent [39]. However, despite favorable survival rates, bladder cancer continues to be the most expensive malignancy as recurrence is extremely common, causing lifelong monitoring of the disease and care to become a necessity [21]. Further complications arise when considering metastatic bladder cancer, for which the 5-year survival rate is merely 5 percent [39]. Thus, metastatic bladder cancer requires quick and personalized treatment that caters to the specific needs of the patient. For this reason, in recent years, immunotherapy has been chosen as a second-line treatment for metastatic cancer, including that of the bladder, and there has been considerable research conducted on the topic.

2.2 Immunotherapy in Cancer Treatment

The advancement of immunotherapy has led to many types of therapies being introduced, including monoclonal antibodies (mAb), cancer vaccines, and adoptive T-cell therapies. However, none of these therapies have proven to be as effective as ICI therapy [73]. ICI therapy is an innovative approach to fighting cancerous cells that functions by suppressing immunological checkpoints such as CTLA-4, PD-1, PD-L1, PD-L2, etc., which cancer cells use to elude the immune system [48]. Otherwise, cancer cells evade immune system detection through the aid of these checkpoints, consequently preventing T cells from eliminating the malignant cells. Multiple studies have been undertaken to develop the most precise ICI treatment since the first ICI drug, Ipilimumab, was approved by the FDA in 2011, denoting a new era in

cancer immunotherapy. Following that, other ICI drugs such as Atezolizumab, Pembrolizumab, and Relatlimab were approved as well [44].
Furthermore, as BC is considered one of the most malignant neoplasms worldwide, many treatment modalities have been introduced, including several different ICI treatments that have been approved for patients with metastatic urothelial carcinoma (mUC), especially for those who are unable to take cisplatin [48]. Correspondingly, a study revealed that around 30 percent of mUC patients responded well to ICI immunotherapy. In addition, around 20-30 percent of urothelial carcinomas expressed PD-L1, indicating a potential biomarker for BC since high levels of PD-L1 expression imply more aggressive tumors, allowing it to be used to predict responses to ICI therapy [48]. Pembrolizumab, the checkpoint inhibitor prescribed to the patients in our dataset, has previously been employed for patients with metastatic urothelial carcinoma and demonstrated significant results [26].

2.3 Challenges in ICI Therapy: Hyperprogressive Disease and the Search for Predictive Biomarkers

However, despite being in the era of advanced immunotherapy for cancer where notable benefits have been observed, the majority of cancer patients, including those suffering from BC, do not receive the intended outcome from ICI treatment. For instance, patients who do not respond to ICI treatment are at risk of developing hyperprogressive disease (HPD), which is characterized by rapidly worsening conditions, and this condition affects 4 percent to 29 percent of patients receiving the ICI treatment. Furthermore, once these patients develop HPD, the risk of death is significantly higher than in patients who responded to therapy. Besides, a study found that HPD is not specific to any one type of cancer [35]. Thus, it is imperative to identify precise biomarkers for the ICI treatment response of BC patients. As a result, researchers are striving to find predictive biomarkers that may assist in estimating how well a patient might respond to ICIs early on in the treatment process [5]. Subsequently, a number of potential biomarkers have recently emerged that may increase the effectiveness of the administration of ICI, including PD-L1 expression, the extracellular matrix, tumor mutation burden (TMB), microsatellite instability (MSI), the microbiome, hypoxia, and interferon-gamma (IFN-) [48]. Nevertheless, the currently accepted biomarkers for ICI treatment do not prove to be strong enough indicators of treatment response [48]. For that reason, it has also been suggested that a composite biomarker, pooling results from the biomarkers mentioned above, promises better results [48].

2.4 Machine Learning and Ensemble Models in Bladder Cancer

Gene expression levels contribute to understanding the underlying mechanisms of a variety of disorders, including cancer [23]. Thus, the dataset GSE111636 is strongly related to this study as it highly correlates with the research objective, i.e. identifying biomarkers for interpreting the difference in reaction to ICI therapy. Further-

more, as the dataset incorporates real-world clinical features, it ensures the model that is being tested reflects actual patient scenarios, making the model clinically efficient.

In cancer research, predictive models built using ML are used to analyze datasets and are trained on clinical and genomic data from patients undergoing ICI treatment. This can be used to assist in creating a more personalized treatment [34]. In recent years, deep learning has also been employed to analyze more complex datasets [41]. However, traditional deep learning models are often regarded as black boxes because they cannot explain how their models reached a certain conclusion, even though they give accurate predictions [57]. This is a severe hindrance in sensitive fields such as healthcare, where results of the intelligent decision-making models require interpretation, sometimes mandated by law. Furthermore, this 'opaque' nature of DNNs may also result in unfairness and bias that escapes the human eye [42]. On the other hand, while single-model approaches using traditional ML algorithms are more easily interpretable, they suffer from algorithmic bias, resulting in different models giving completely different outputs, and a lack of generalizability, especially when working with small sample, high-dimensional datasets [66].

Ensemble models based on traditional ML models solve both these problems as they incorporate the interpretability of simple ML models while also working well when there is high disagreement between the individual models, reducing errors caused by bias and variance. This is evidenced by their usage on heterogeneous datasets from cancer patients involving genetic variations, where they gave more stable and generalized predictions as compared to single models [71], and their effectiveness in high-dimensional data environments, such as cancer gene expression datasets [65].Furthermore, ensemble modeling has also proven to be more accurate than singular machine learning algorithms for tasks such as cancer classification [5] [68]. However, there has yet to be research done on the role of ML, and consequently, ensemble modeling in identifying predictive biomarkers in BC. Finally, although deep neural networks have shown greater capabilities when it comes to classification and feature selection when compared to traditional machine learning algorithms, no literature that compares ensemble models with deep neural networks exists [54].

To aid in bias reduction and enhance the generalizability of the models used in the research, nested LOOCV with Stratified K-fold cross-validation is incorporated. Although several other cross-validation techniques exist to estimate generalization performance, this approach with LOOCV is used as it caters to small datasets; in LOOCV, since the dataset is trained on nearly the entire dataset in each fold, it results in the greatest reduction of bias and provides better generalizability keeping in mind the limitations of the dataset, with the primary tradeoffs being increased computational cost and higher variance [68].

Furthermore, while the results of ensemble models can be aggregated in multiple ways, such as bagging, boosting, stacking, this study uses a frequency-based majority voting aggregation technique, which has shown better accuracy when compared to individual models [25]. While bagging, boosting, and stacking can provide better results, they have their own caveats. For instance, the effectiveness of boosting is dependent on the dataset, as it is prone to overfitting when working on noisy and small datasets [3]. Moreover, bagging also risks overfitting to the dataset if base learners become too complex, which is more of a possibility when working with small datasets [28]. Alternatively, despite stacking providing better performance,

it requires intricate implementation and suffers from a lack of interpretability [61]. Thus, the simplicity and ease of interpretability of frequency-based majority voting made it the primary choice for our ensemble model. By tracking how often a feature is selected across multiple folds or models, this method emphasizes consensus among all the models and filters out model-specific noise, selecting features that have a high consensus stability across various different models and resulting in a more robust feature set.

2.5 Key Findings and Research Gap

BC is a common health issue occurring mostly in adults, and men are more likely to be affected by this [39]. It remains the most costly cancer due to its high recurrence rate, and poor prognosis in metastatic BC increases the cost even further [21]. Hence, researchers have shifted to other treatment modalities such as immunotherapy to expedite the treatment process. Immunotherapy, such as ICI treatment, is recognized as a revolutionary treatment for cancer and regarded as a fine-line treatment option, especially for patients who are unable to take cisplatin, a type of chemotherapy [51]. ICI treatment, such as pembrolizumab (PD-1 blockade), works by blocking the checkpoints that cancer cells use to evade the immune system. However, despite ICI therapy being a new and effective treatment option, it has some disadvantages that need to be addressed.

One of the major concerns of ICI therapy is the risk of developing HPD, where the cancer grows much faster after ICI treatment [35]. As a result, if doctors are unable to predict whether a certain patient will respond to ICI treatment, patients run the risk of developing HPD. Patients are also at risk of incurring unnecessary financial burden if treatment is ineffective as ICI therapy can cause upwards of US $100,000 per year of treatment. Therefore, the need for reliable biomarkers is essential to help predict which patients will benefit from ICI treatment. Although there are certain biomarkers present, such as PD-L1 expression, tumor mutation burden (TMB), and microsatellite instability (MSI) [49], these are not always reliable as independent biomarkers. This urges the need for a range of different biomarkers to predict treatment response as early and accurately as possible. On that account, this research addresses this gap by finding genes that are indicative of response to ICI treatment, specifically through the administration of Pembrolizumab, and can function as biomarkers for treatment response.

Furthermore, ML methods are widely recognized in tasks such as cancer classification, biomarker identification, and even response to ICI treatment. Among machine learning techniques, Deep Neural Networks (DNNs) have demonstrated high accuracy; however, they require a large dataset for training, which also increases computational cost [41], while also being expensive to create and obtain. Due to this high cost, gene expression datasets are most often high-dimensional with a limited sample size, which makes deep learning methods prone to overfitting. Besides, these state-of-the-art deep neural methods are regarded as the “black boxes” because their inner workings are hard to understand, lacking the transparency that is required in healthcare applications [57]. On the other hand, single models based on traditional ML models lack in generalizability and suffer from algorithmic bias when working with such small datasets [67].Thus, there is a void in research regarding the identification of biomarkers of ICI therapy response in BC patients through compu-

tational means. To address these issues, our study uses ensemble feature selection technique, coupled with cross-validation and frequency-based majority voting. This robust technique enables an improved feature selection procedure, selecting the most significant genes.

Chapter 3

Impact Analysis and Design

3.1 Final Specifications and Requirements

Our proposed plan is to design an ensemble model that aims to identify and validate potential biomarkers that can be used to predict the efficacy of ICI in BC patients. The dataset used in this study is GSE111636 from the Gene Expression Omnibus (GEO). It consists of gene expression measurements of 70,523 genes obtained from formalin-fixed paraffin-embedded (FFPE) tissue samples collected from 11 BC patients.
The primary task involves biomarker identification from the binary classification where patients are categorized into two distinct classes based on their response to ICI therapy; class 1 corresponds to responders, referring to patients who exhibited a positive response to the treatment, while class 0 corresponds to progressors, or patients whose condition worsened during the course of therapy.
The goal is to identify predictive biomarkers - specific genes that significantly influence the classification decision made by the ensemble learning model. These biomarkers can serve as crucial indicators in understanding patient-specific responses to ICI treatment and guiding personalized therapy strategies.

3.2 Societal Impact

This research aims to advance the realm of personalized treatment. By utilizing real-world clinical data and an ensemble machine learning model, the study enables clinicians to tailor immune checkpoint inhibitor (ICI) therapy to individual BC patients according to the identified biomarkers, reducing their reliance on generalized, trial-and-error approaches. Since the dataset reflects actual clinical conditions, the model developed becomes not only technically sound but also clinically applicable—ensuring more efficient, targeted, and relevant treatment strategies.
Another major aim of this research is the reduction of side effects and improvement of survival rates. Early identification of patients who are unlikely to respond to ICI therapy using the help of biomarkers helps clinicians adjust or redirect treatment plans more appropriately and prevent the onset of HPD.

3.3 Environmental Impact:

By minimizing the trial and error phase, our research can reduce the number of hospital visits, mitigating unnecessary imaging scans and drug administration. This lowers the environmental footprint associated with cancer care.

3.4 Ethical and Legal Considerations:

Ethical considerations and legal compliance are central in the development of any machine learning models in the scope of healthcare. One major concern is bias and fairness as ML models can be affected by biases because of imbalanced or non-representative datasets, which directly results in disparities in predictions across certain demographic groups. This can be tackled by incorporating model evaluation strategies by our ensemble model. This issue is addressed in our approach by incorporating robust model evaluation strategies within the ensemble framework to detect and mitigate bias. Another critical aspec of biomedical research is data privacy. Our model uses anonymized patient data sourced from one of the most reputable public repositories for bioinformatics data, which is NCBI's Gene Expression Omnibus (GEO) database. Using this, it is ensured that there is no violation of privacy rights under data protection regulations such as HIPAA or GDPR in the clinical development phase.The dataset used in this research also adheres to MIAME standards, enhancing the clarity and usability of the dataset for research and clinical application. In addition, model interpretability is essential in fields such as biomedicine in order to foster trust and ensure accountability in clinical settings. To address this, the model proposed in this research uses simple and interpretable learners and aggregation mechanisms.

3.5 Economic Considerations

Our research aims to reduce healthcare cost by identifying biomarkers that are indicative of treatment response in BC patients undergoing ICI therapy. By accurately identifying non-responders early, hospitals can avoid administering costly treatments that are unlikely to be effective, thereby saving valuable medical resources and reducing the financial burden on both healthcare systems and patients. Moreover, personalized treatment strategies, made possible through predictive modeling, can shorten the overall treatment lifecycle. This not only enhances therapeutic efficiency but also reduces hospitalization durations and associated costs, leading to more sustainable and cost-effective healthcare delivery.

Chapter 4

Implementation and Methodology

4.1 Data Collection and Preprocessing

4.1.1 Description of the Dataset

The rapid increase in transcriptome profiling tools has increased the availability of gene expression datasets. One of the most reputable public repositories for bioinformatics data is NCBI's Gene Expression Omnibus (GEO) database [32]. This study utilizes GSE111636 dataset [29], titled, "Identification of potential biomarkers of response and resistance to programmed cell death-1 blockade in patients with advanced urothelial tumors," obtained from the GEO database. The dataset comprises formalin-fixed, paraffin-embedded (FFPE) bladder tissue samples of patients undergoing immunotherapy with pembrolizumab (MK-3475), which is an anti-PD-1 immune checkpoint inhibitor (ICI) used to treat patients with advanced urothelial carcinoma. RNA samples were profiled using the Affymetrix GeneChip Human Transcriptome Array 2.0 (HTA2.0) platform (GPL17586), Furthermore, GEO stores gene expression data by following MIAME guidelines, ensuring interpretability and shareability of the data, making this data source suitable for our study [7]. Alongside source reputation, the choice of this dataset is influenced by several factors, including accessibility of data and alignment of the data with our research. Specifically, the GSE111636 dataset focuses solely on BC patients with advanced urothelial tumors, whose DNA samples were analyzed using the GeneChip Affymetrix Human Transcriptome Array 2.0 platform, measuring the expression levels of genes across samples. These patients were treated with pembrolizumab (MK-3475), which is a programmed cell death-1 (PD-1) blockade and the dataset encompasses comparisons between those who responded positively to the treatment and those who reported disease progression, categorized into responders and progressors respectively. The primary aim of this dataset is to identify probable biomarkers associated with response and resistance to PD-1 blockade immunotherapy.

4.1.2 Dataset Size

The profiling of a total of 11 RNA samples allowed for the high-resolution detection of over 70,523 transcript clusters. The samples in the dataset were classified based on their clinical response to therapy as either responders (responding positively to ICI therapy) or progressors (not showing positive response to ICI therapy); out of the 11

samples, 6 were labeled as responders and 5 as progressors. This binary classification provided a robust foundation on which to perform differential expression analysis, biomarker identification, and pathway enrichment research.

Before Preprocessing

Initially, the gene expression dataset consisted of 70,523 rows and 11 columns, where each column represented the individual ID of each sample patient and each row specified a unique gene probe or Affymentrix HTA2.0 Probe ID a specific gene; the values in rows showed the activity or gene expression values of that specific gene in a particular patient. However, the raw dataset also contained irrelevant data that reduces the accuracy of research, making preprocessing essential for meaningful analysis.

After Preprocessing

The dataset used in this study underwent several preprocessing steps to ensure quality and consistency. The first step of preprocessing involved transposing the dataset. Since the features for our dataset were the Probe IDs for the genes and their respective expression values, these values had to be shifted to columns. Consequently, the sample IDs had to be shifted to rows. This was a necessary step in order to make the dataset function properly with Python's built-in functions as these functions expect datasets to have samples in rows and features in columns. Furthermore, a separate column for the target variable that served as a label for responders and progressors was added according to the dataset description where each sample ID was clearly identified as a responder or progressor; the target variable denoted responders as "1" and progressors as "0".

The next step of preprocessing involved dealing with missing and duplicate values in the dataset. To address this, columns that had missing values were removed, bringing down the total number of features from 70,523 to 30,266. Afterwards, data duplication was checked but no duplicate data points were found, leaving the dataset unchanged.

However, a low noise to signal ratio (SNR) is characteristic of gene expression datasets and such was the case for this dataset as well. This has a detrimental effect on downstream analyses, leading to higher false discovery rates and missed true signals [12]. The effects are even stronger for small datasets, resulting in inconsistent and poor reproducibility of outputs [13]. Thus, as a means to increase the SNR of the dataset, percentile-based variance thresholding was applied. This resulted in a reduced dimensionality of the dataset as features that did not have a certain amount variance, and consequently were not differentially expressed, were discarded. Percentile-based variance thresholding has been shown to positively affect downstream analyses, effectively controlling false discovery rates [8]. The value for our percentile based thresholding was to 95 percent, i.e. only the features that had a variance of within the top 5 percent were kept. Although lower percentages were considered, they consistently resulted in overfit outputs and thus, 95 percent was chosen as it provided a good balance between acceptable results while also not setting the variance threshold too high, risking discarding genes that may be differentially expressed. Therefore, at 95 percent, the variance threshold value was 0.1918

and the resulting dataset had 1514 features; the minimum variance in the dataset then became 0.0002 and the maximum variance 24.5168.

4.1.3 Feature and Target Variable

The primary features of this dataset, organized in columns, comprised gene expression values, representing the transcription levels of various genomic data across the patients, with the specific Affymetrix HTA 2.0 Probe IDs for the genes being the column headers. The target variable in this study was the patient response to ICI therapy, which was categorically defined into two separate classes: (i) responders, who were patients that showed a positive response to the treatment, denoted by 1, and (ii) progressors, who were patients whose conditions had worsened over the duration of treatment despite undergoing the therapy, denoted by 0. This classification served as the output or target variable.

4.2 Model Design and Description

4.2.1 Description of the models used

This research involves a careful selection of various models based on their strengths, specifically in terms of feature selection. These models include SVM-RFE, RF-RFE, LR-RFE, and MI alongside two deep learning models, namely Concrete Autoencoder and a Deep Neural Network with L1 Regularized Logistic Regression.
For SVM-RFE, the linear kernel is used , bringing the number of linear models and non-linear models to 2 each as Logistic Regression is also a linear model while Random Forest and Mutual Information are non-linear models. The reason for choosing equal numbers of linear and non-linear models is to balance the feature selection outputs as both types of models have shown to perform better in certain conditions [37]. It must also be mentioned that the selection threshold for each model is selected in a way that results in around 30 features being selected. In addition, to ensure uniformity across all gene expression features and eliminate scale-induced bias during model training, the dataset was standardized using z-score normalization, which centers the data around a mean of zero with a standard deviation of one, allowing each gene to contribute equally to the learning process.
Furthermore, all of the models have nested cross-validation as well as LOOCV incorporated into them as they both reduce bias in error estimates and performance and provide the most robust feature selection outcomes [30] [69]. While LOOCV is applied to the outer loop of the nested cross-validation, the inner loop is used to tune the hyperparameters for the individual models - stratified K-fold cross validation is used in the inner loop to ensure consistent and reliable tuning. In the end, the final model with the best selected hyperparameters is tested on the outer loop. In addition, special care is taken to prevent data leakage by ensuring that models only learn from the training data that is made available to them. As a result, the models do not have access to the test data beforehand, resulting in the outcome not being biased by the test data.
The model implementation for all the models is intentionally kept constant in order to establish a fair comparison, ensuring that the difference in the results could clearly and easily be attributed to only algorithmic differences.

SVM-RFE

Support Vector Machine Recursive Feature Elimination (SVM-RFE) is a widely used wrapper-based feature selection algorithm that operates by recursively training a support vector machine and eliminating features that contribute least to the decision boundary. To be more specific, SVM-RFE begins by fitting a linear SVM model to the training dataset and ranking all features based on the magnitude of their weights in the model's decision function. Features with the smallest absolute weights, indicating the least contribution to classification, are iteratively removed from the dataset, and this process is repeated until a specified number of top-ranked features remain [4]. A key benefit of SVM-RFE is that it considers interactions between features in its selection process, making it particularly effective in high-dimensional biomedical datasets such as ours where feature interdependence can influence patient outcomes. SVM-RFE has consistently shown to provide better results than other feature selection algorithms for functions such as cancer classification, making it a good fit for this research [4].

To integrate this algorithm into our study, we implemented a complete pipeline that combines SVM-RFE with nested cross-validation using LOOCV on the outer loop and stratified K-fold validation on the inner loop for hyperparameter optimization. The first step involved normalizing the dataset with z-score scaling to ensure all gene expression values are on a consistent scale. Afterwards, in each outer LOOCV fold, SVM-RFE is used to reduce the dimensionality of the data by selecting the top 2 percent of features, after which these features are passed into a GridSearchCV routine. This routine tunes the hyperparameters C and gamma for a final non-linear SVM classifier with an RBF kernel using 3-fold stratified inner validation. After the optimal model is selected, it is then evaluated on the left-out test sample from the outer fold. This approach, thereby preventing data leakage and maintaining the integrity of model evaluation.

Random Forest

Random Forest Recursive Feature Elimination (RF-RFE) is a variant of the traditional RFE algorithm that utilizes the feature importance scores derived from a Random Forest classifier, which has shown great results in gene expression datasets [6], to iteratively eliminate less relevant features. In contrast to SVM-RFE, which ranks features based on model weights, RF-RFE evaluates the contribution of each feature to classification accuracy using impurity-based metrics. Then, at each step, features contributing the least to reducing classification error are discarded, and this process is repeated until a predefined number of top features is selected. Random Forest is inherently non-linear and capable of capturing complex gene-gene interactions, making it particularly useful for high-dimensional datasets like ours where underlying relationships may not be strictly linear, a trait that is not uncommon in gene expression datasets.

In this study, RF-RFE was applied through a standardized pipeline that is similar in structure to the SVM-RFE model for consistency. Initially, gene expression values are standardized using z-score normalization. LOOCV is then employed for robust outer loop evaluation, with RF-RFE performed in each outer fold. This resulted in the top 2 percent genes being retained in every iteration based on their importance scores. A Random Forest classifier with 100 estimators is then trained using only

the selected features and evaluated on the left-out sample.

Logistic Regression

Logistic Regression Recursive Feature Elimination (LR-RFE) is also a wrapper-based feature selection method that utilizes weight coefficients, generated by L1-regularized logistic regression, to discard features based on their informativity. LR-RFE evaluates feature relevance directly through the magnitude of these coefficients, with L1 regularization providing a penalty term that effectively reduces coefficients of unimportant features to 0 [1]. In this context, features with the smallest absolute coefficient values are iteratively removed until a predefined number of top-ranked features remain. The simplicity of LR-RFE makes it especially suitable for high-dimensional data as it tends to retain only the most discriminative features while still being interpretable, a characteristic that is invaluable for work in such biomedicine. Similar to the implementation of SVM-RFE and RF-RFE, LR-RFE was implemented through a robust and transparent pipeline that integrates this feature selection approach with LOOCV in the outer fold and Stratified K-fold CV in each inner fold for unbiased performance evaluation. The dataset is first standardized using z-score scaling, ensuring gene expressions were not disproportionately influenced by scale differences. During each LOOCV fold, LR-RFE is applied exclusively to the training set using a logistic regression model with an L1 penalty to identify the top 2 percent of features based on their coefficient magnitudes. The selected features are then used to train a final logistic regression classifier in each inner fold, which is subsequently evaluated on the left-out test instance in that fold. Hyperparameter tuning is also carried out in each inner fold using a GridSearchCV routine; the parameters include the type of penalty, which may either be L1 or L2, and the value of C, which controls the regularization strength. This strategy ensures a clear separation between training and test data within each iteration, effectively minimizing data leakage and preserving the validity of the performance estimates.

Mutual Information

Mutual Information (MI) is a filter-based feature selection technique that quantifies the dependency between each feature and the target variable using "mutual information", a non-parametric measure rooted in information theory, which essentially quantifies how much knowledge about a certain feature reduces uncertainty about another. Unlike correlation coefficients that only capture linear relationships, mutual information detects both linear and non-linear dependencies, making it particularly suitable for high-dimensional biomedical datasets where interactions cannot strictly be separated into linear and non-linear interactions. MI's effectiveness is evident as it has consistently shown better outcomes than other screening methods [31].
Much like the previous models, MI based feature selection was integrated into our research through a rigorous classification pipeline tailored for high-dimensional gene expression analysis. To ensure robust model evaluation and guard against overfitting, we coupled MI-FS with LOOCV in each outer loop and Stratified K-fold CV in each inner loop. Within each outer LOOCV fold, the training data undergoes feature scaling via z-score scaling, after which mutual information scores are computed between each gene feature and the binary target. Afterwards, the top 2 percent

of features with the highest scores are retained, reducing the dimensionality of the dataset.
Moreover, in each inner loop, a Logistic Regression classifier is trained on the reduced feature set, with hyperparameters optimized via running GridSearchCV using stratified 3-fold cross-validation. This nested design maintains strict separation between training and testing data, thereby minimizing information leakage and preserving the validity of performance estimates.

Ensemble Modeling

Ensemble feature selection is a robust method that integrates the outputs of multiple base feature selectors to identify a consensus subset of informative features. Rather than relying on the selection bias or methodological assumptions of a single model, ensemble methods instead aggregate across diverse selectors to capture a wider range of relevant signals, thus enhancing both stability and generalizability by essentially harnessing the unique strengths of each model. However, the success of an ensemble model is determined by the effectiveness with which the basic learners are taught and integrated [58].
As mentioned before, we applied four distinct feature selection strategies - SVM-RFE, RF-RFE, LR-RFE, and MI - within a nested LOOCV framework. For each model, feature selection was independently executed in every outer fold, yielding a set of around 30 top-ranked features per fold based on the specific selection criterion This produced four collections of fold-wise feature subsets, each representing the model-specific perspective on relevant features.
However, to perform ensemble feature selection, a frequency-based majority voting mechanism was introduced that aggregates the fold-level outputs for each of the four models. Specifically, the features selected in each outer fold for each model were compiled, computing a selection frequency score for every feature based on how often it was chosen across the folds. Features that appeared consistently - i.e., selected in the majority of LOOCV iterations across models - were considered more likely to be robust and biologically meaningful.
This ensemble methodology offers several advantages in the context of high-dimensional data as it leverages the strengths of diverse algorithms while also improving the stability of feature selection.

Deep neural networks

For gene expression data, which typically has many features but relatively few samples, traditional models often outperform deep learning models unless applied intricately. Thus ,to provide a fair comparison, this research also applies two deep learning models, namely an Autoencoder and a Deep Neural Network based on L1 Regularization, to carry out feature selection as well.

Autoencoders

Gene expression datasets often contain thousands of genes, resulting in high-dimensional data that can be challenging for traditional models to handle effectively. Autoencoders are particularly suitable in this context because they can compress the data

into a lower-dimensional latent space representation, preserving the essential patterns while reducing dimensionality. This compression facilitates more efficient feature learning by enabling the model to focus on compact and informative features, often improving performance without significant loss of information. Additionally, autoencoders can help remove noise and measurement errors commonly present in gene expression data. The autoencoder architecture consists of two parts: the encoder, which maps the input to a latent representation, and the decoder, which attempts to reconstruct the original input from this compressed encoding. The model is trained by minimizing the reconstruction loss, which encourages it to learn only the most essential features. In this research, these latent features were extracted from the encoder and used as inputs to classifiers such as Logistic Regression, Neural Networks, and XGBoost for prediction tasks. However, although autoencoders can model complex gene interactions and are unsupervised - an advantage when annotated data is scarce - they have certain caveats. For instance, compressed latent features do not always correspond directly to specific genes, which can reduce biological interpretability. Furthermore, there is always a risk of overfitting, especially when dealing with deeper or more complex variants like variational autoencoders.

DNN with L1 Regularization

Complementing this, a Neural Network with L1 regularization is implemented as well, which is also particularly useful for high-dimensional biological data. Harnessing the power of L1 regularization as mentioned before results in a more interpretable model that is less prone to overfitting, effectively improving generalization to unseen data—an important consideration when sample sizes are limited. However, while L1 regularization simplifies the network and focuses it on the most predictive genes, it can sometimes prune too aggressively, potentially discarding useful information when multiple features share similar signals. This may end up causing training to become unstable. Nonetheless, this approach enhances robustness and interpretability, making it well suited for gene expression classification tasks where feature selection is priority.

The overall outline of the research is shown in Figure 4.1.

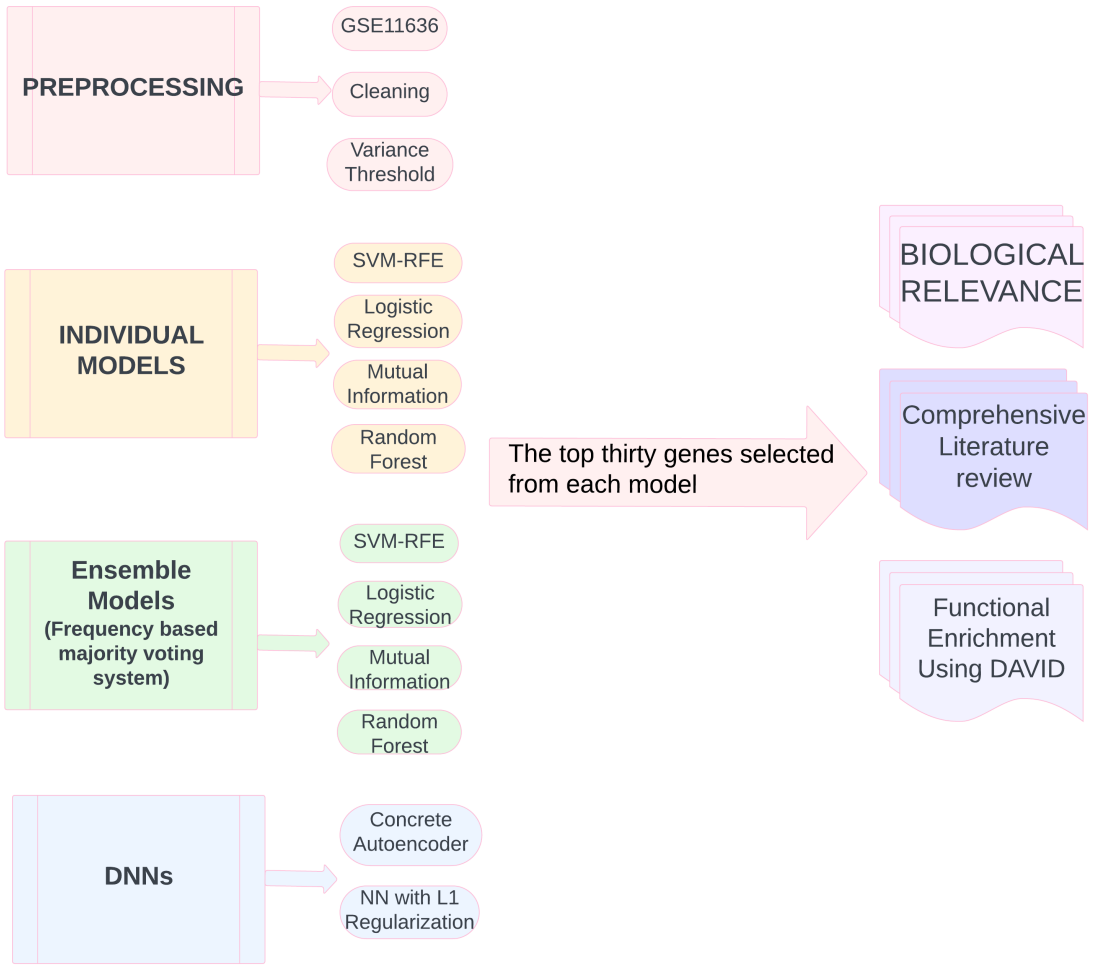


Figure 4.1: Research Outline

4.3 Model Training and Evaluation Setup

4.3.1 Performance Evaluation Metrics

Jaccard Similarity

To evaluate the stability and reliability of the selected features, the top features in each fold were tracked, computing the mean Jaccard similarity coefficient across all pairs of LOOCV iterations - this was done for each model separately. The use of Jaccard similarity quantifies the consistency of feature subsets and provides insight into the robustness of feature selection under varying training conditions.The mean Jaccard Similarity was calculated using Formula (4.1).

Jaccard Formula

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{4.1}$$

Spearman Correlation Coefficient

Additionally, the mean Spearman rank correlation coefficient was also calculated using Formula (4.2) to assess agreement in feature importance rankings. The mean Spearman rank correlation coefficient provides insight into the feature selection consistency and reproducibility, making it a common metric in biomedical datasets. This dual perspective—subset overlap and score consistency—offers a nuanced view of feature selection stability, a critical aspect when interpreting biomarkers for downstream biological validation.

Spearman Correlation Coefficient Formula

$$\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} \tag{4.2}$$

LOOCV Accuracy

Finally, mean LOOCV accuracy was calculated across all models using Formula (4.3). This is a fairly simple performance metric, essentially finding the average LOOCV accuracy across all the folds.

LOOCV Formula

$$\text{LOOCV error} = \frac{1}{n} \sum_{i=1}^{n} L\big(y_i, \hat{f}_{-i}(x_i)\big) \tag{4.3}$$

Method	**Metric**	
SVM-RFE	LOOCV	0.6364 ± 0.4810
	Jaccard	0.3493
	Spearman-correlation	0.6895
RF-RFE	LOOCV	0.5455 ± 0.4979
	Jaccard	0.2004
	Spearman-correlation	0.3409
MI	LOOCV	0.3636 ± 0.4810
	Jaccard	0.2582 ± 0.0861
	Spearman-correlation	0.5256
LR-RFE	LOOCV	0.6364 ± 0.4810
	Jaccard	0.4004
	Spearman-correlation	0.6863

Table 4.1: Results and Statistical Significance

It is clear from Table 4.1 that SVM-RFE and LR-RFE perform better than the other models in terms of LOOCV accuracy and Spearman Rank Correlation Coefficient, followed by RF-RFE and MI. However, while RF-RFE has better LOOCV accuracy,

MI has a better mean Jaccard similarity and Spearman Rank Correlation Coefficient. This indicates that while using RF-RFE results in feature sets that give slightly higher predictive accuracy, using MI results in feature sets that are more consistent and stable. Nevertheless, this is an indication of the algorithmic bias of different models; our ensemble model essentially aims to harness the different strengths of each model, effectively pooling their results for the best outcome.

In addition, it is imperative to mention that these values should not be seen as an end-all-be-all metric through which to judge the performance of these models on such a limited dataset and should be observed intricately. For instance, although a mean LOOCV accuracy of 0.3636 to 0.6364 may seem low, the conditions under which these predictions were made should also be considered. The classifiers for all the models essentially made their predictions based on the feature-selected dataset alone, choosing 5 out of 30 features. Thus, the classification did not consider biological relevance or correlations between the features, finally making "biologically blind" predictions, resulting in low accuracy. Nevertheless, 3 out of 4 models still predicted at better-than-chance levels. Another point of contention could be the high variance in mean LOOCV accuracy values; however, this is merely a characteristic of LOOCV. Since the model is tested on a single data point, the accuracy can either be either 1 or 0, explaining the high variance.

Furthermore, although the mean Jaccard similarity values may be considered low at first glance, their results should be viewed in context of the dataset. Since, the feature selection process selected around 30 features from 1514 features in total, most of which were not differentially expressed or relevant and merely added noise, a mean Jaccard similarity value of 0.2004 to 0.4004 may be acceptable. A mean Jaccard similarity value of 0.2004 to 0.4004 essentially means that 20 percent to 40 percent of the 30 features selected (6 to 12 features) were consistently selected among all folds.

Tying in to the mean Jaccard similarity value, although the mean Spearman Rank Correlation Coefficient may be considered low at first sight, it is merely a consequence of the Jaccard value. Since around 6 to 12 features were consistently selected across all the folds in the first, even if their rankings in each fold differed even slightly, it would lead to low Spearman Rank Correlation Coefficient values.

P-value as a performance was also considered but finally not implemented due to various reasons even though for over a decade, P-values have been one of the commonly used tools for statistical analysis. The probability p-value judges whether an observed result is likely due to chance, and a value less than 0.05 is considered statistically significant. Although this threshold can be considered in most cases, for a small sample size, bias, or random error can make the p-value unreliable [19]. This is primarily observed when interacting with genomic data, which typically has a small sample size. As a result, we had to look beyond the common p-value for biological systems, as the relationship can be intense in this context.

Thus, the intricate gene-gene interaction and the unique constraints of our study made it imperative to understand both the statistical significance and biological relevance of our potential biomarkers. This was especially necessary for our study as we incorporated gene-gene interactions, where a significant statistical result might sometimes be meaningless due to a lack of biological context and vice versa. Thus, our study also put considerable emphasis on the individual gene interactions, pathway analyses, and valid literature to connect all the dots and support our results.

Besides, this comprehensive analysis also ensures that our findings are meaningful in a biological context.

4.4 Biomarker Identification

Evidently, incorporating biological relevance into the biomarker identification process was inevitable. However, before carrying that out, narrowing down the feature-selected dataset was crucial as studying every single gene identified by every model would result in a tedious and inefficient analysis of around 300 genes. Therefore, another layer of frequency-based weighted majority voting was adopted, where all genes finally selected by all models were considered. However, genes selected by ensemble models were given higher weight as these genes were consistently selected across many folds, meaning that it was unlikely that these genes were chosen merely by chance. On the contrary, since only the last fold feature sets of the single models were considered, the chances of any gene being picked up by chance was considerably high. The same was the case for the deep learning models. Thus, this voting system was chosen.
The final feature set after the frequency-based weighted majority voting process resulted in a ranking that put genes selected by all models first, followed by genes that were selected by ensemble models and single models or ensemble models and DNNs, and genes selected by ensemble models only; genes that were only selected by 1 single model or DNN was discarded due to the random chance factor mentioned above. The final check for biological relevance was carried out on this ranked feature set.

Chapter 5

Biological Validation and Model Interpretability

5.1 Biological Validation and Functional Characterization

Our robust feature selection method successfully identified 11 genes collectively using ensembles and individual techniques. Given the fact that working with a small patient cohort (N=11) has intrinsic limitations that make it challenging to prove direct statistical significance. Therefore, it is crucial to show the biological relevance of these 14 identified genes. To provide clinical validation, this section focuses on describing the functional role of these genes and connecting them to the underlying mechanism of BC, ICI therapy, and response to ICI.

5.1.1 Gene-Specific Biological Relevance:

The gene set determined by our model plays interrelated roles in cancers, immune regulation, and the effect on ICI treatments. For instance, the nuclear enzyme TOP2A has been identified by both ensemble and single models, and is highly relevant in BC. Because a higher amount of TOP2A indicates a more aggressive cancer. Recent studies also showcase that patients with a high level of TOP2A have less chance of surviving [33]. Hence, blocking TOP2A reduced dissemination of cancer, and thus identified as a prognostic biomarker for BC. Further investigation presented that TOP2A increases PD-L1, which helps cancerous cells evade the immune system. Therefore, this implies that it is best to consider the TOP2A gene while administering ICI therapy.
One of our crucial genes identified by our ensemble model is ZEB2, which is associated with poor prognosis. To further elaborate, ZEB2 has proven to protect BC cells from DNA damage-induced apoptosis [56]. As a result, it promotes tumor development. Aside from cancer progression, ZEB2 also plays a pivotal role in the immune system. Because it is involved in a process known as epithelial–mesenchymal transition (EMT) [56]. This course of action faciliates cancerous cells to elude from the immune system while contributing them to become more aggressive. Studies also reveal that when ZEB2 is highly expressed, it contributes to the progression of tumor environment immunosuppressive [47]. In conjugation with this, ZEB2, in combination with other genes in a gene signature, was highly correlated with immune

checkpoint inhibitors, such as CTLA-4, PD-1, PD-L1, and PD-L2 [47]. This yields in treatment resistance to drugs like pembrolizumab, which makes ICI treatment and other similar therapies ineffective.
Furthermore, genes, such as PTCH1, are involved in the Hedgehog (HH) signaling pathway, and this potential biomarker was identified by both ensemble and single models. This gene exerts a vital function in the (HH) pathway because it acts as a key regulator that keeps the HH signaling turned off when there is no HH ligand present [76]. A contemporary study found that HH signaling plays a pivotal role in BC because the malignant cells from the bladder lining manifested continuous HH signaling, and this HH activity was imperative for the cancer cells to mature [11]. Hence, the (HH) pathway can be blocked to lessen PTCH1 volume. To the contrary, a further analysis demonstrated that a mutation in the PTCH1 gene permits an enhanced response to ICI treatment in gastrointestinal and colorectal cancers [55] [59].
Beyond that, the cholesterol regulator or HMGCR gene (identified by both models), reveals that the rate of HMGCR greatly influences cancer progression owing that higher HMGCR levels are linked to the growth and deterioration of BC [72]. Drugs that are employed to lower cholesterol levels, such as HMG CoA reductase (HMGCR) inhibitors, commonly known as statins, may help aid in BC [72] [74]. To further justify this, atorvastatin (ATO), when treated with ICI therapy, increased the survival rate in lung cancer patients [50]. This denotes PTCH1 and HMGCR genes ought to be treated accordingly before administering ICI treatment.
Another study aimed to see the level of UPK1B, a protein that regulates permeability in bladder epithelial cells. UPK1B promotes cancer cell growth and invasion by activating the Wnt/β-catenin pathway [27]. To reinforce the reasoning, a high level of UPK1B is often found in more advanced tumor stages, which suggests that BC has spread to the lymph nodes and other parts of the body, resulting in poor prognosis. This potential biomarker could only be identified by our ensemble model.
Consequently, the protein that functions in DNA repair, CDCA5, identified by both mechanisms, promotes the growth of BC cells. This gene is upregulated in BC, which implies that patients with high levels of CDCA5 have adverse survival outcomes. The tumor promoting role of CDCA5 is validated through the augmentation of specific proteins called CDC2 and Cyclin B1, and by stimulating the PI3K/AKT/mTOR pathway [36]. In addition, CDCA5 increases PD-L1 levels on cancer cells; hence, blocking both PD-L1 and CDCA5 terminated lung cancer cells from growing [64]
Other genes that contribute to the tumor microenvironment are known as CD248 (identified by both ensemble and single). Interestingly, these are predominantly observed in the blood vessels around the tumor, and this biomarker also plays a key role in the growth and development of BC cells. Again, patients with a high level of CD248 experienced poor prognosis, as it is defined as an upregulated biomarker [46]. It is also suggested that a treatment that inhibits both PD-1 and CD248 performs better against cancer, especially for individuals for whom merely blocking PD-1 is ineffective [62].
Like many others, ACTL6A (identified by both ensemble and single) has been identified as a key determinant, with high expression associated with poor prognosis among various malignancies [63].
Similarly, elevated levels of NORAD (identified by both ensemble and single) resulted in limited survival rate, and lowering them improves survival ratios [24].

Another study indicates that S100P (identified by both ensemble and single) is often used as a marker when clinicians are unclear of the origin of a high-grade tumor [22]. There is no clear evidence that shows S100P is responsible for the growth and development of BC. but they constitute key takeaway in distinguishing where the tumor is originated from when diagnoses are challenging. However, from this we can infer that S100P is highly expressed in aggressive urothelial carcinoma. Besides, studies also showcase that S100P is not limited to BC and can be found in other cancers as well [22].
Lastly, our ensemble model also identified another strongly overexpressed gene called TAGLN2. Studies demonstrated that TAGLN2 is a non-invasive biomarker for BC detection, and it can further examine if cancer has spread to lymph nodes [15].

Gene Name	**Symbol**	**Functionality**
DNA topoisomerase II alpha	TOP2A	A nuclear enzyme that controls the shape and structure of DNA during replication and transcription.
Zinc finger E-box binding homeobox 2	ZEB2	DNA-binding transcriptional repressor that operates inside the nucleus and interacts with SMAD protein.
Cell division cycle associated 5	CDCA5	A protein that functions in DNA repair and chromatin attachment.
Patched 1	PTCH1	A receptor protein that is involved in controlling the hedgehog signaling pathway.
Actin-like 6A	ACTL6A	A gene that produces an actin-related protein which is part of the BAF complex and activates certain genes by opening up condensed DNA.
Transgelin 2	TAGLN2	Produces a transgeling-related protein that can accurately detect tumors and may help in determining if cancer has spread to lymph nodes.
3-hydroxy-3-methylglutaryl-CoA reductase	HMGCR	An enzyme that regulates the body's cholesterol production and is targeted by a drug to reduce cholesterol levels.
S100 calcium binding protein P	S100P	A calcium-binding protein which is linked in cell growth and differentiation.
CD248 molecule	CD248	A protein-coding gene which is linked to lymph node development, blood vessel cell death, cell movement, it may also help cells to bind to extracellular matrix.
Uroplakin 1B	UPK1B	A cell surface protein that may help regulate membrane stability and permeability in bladder epithelial cells, mainly found in bladder cancer.
Non-coding RNA activated by DNA damage	NORAD	A long non-coding RNA activated by DNA damage and aids in regulating cell stability.

Table 5.1: Overview of the identified genes

Table 5.1 highlights the identified genes along with their gene symbols and functionalities.

5.1.2 Functional Enrichment Analysis

To gain a holistic understanding of the biological relevance of the identified genes, we conducted functional enrichment using the DAVID Bioinformatics Resource. This enrichment enabled us to understand their combined effects while complementing individual gene annotation. It also gives us a more in-depth understanding of their biological significance, particularly in the context of BC and the outcome for ICI treatment.

David Enrichment Analysis

Functional enrichment analysis was performed utilizing the Database (DAVID) Bioinformatics Resources, and 11 potential biomarkers (TOP2A, ZEB2, PTCH1, HMGCR, UPK1B, CDCA5, CD248, ACTL6A, NORAD, S100P, TAGLN2) were given as input. Table 5.2 depicts the result derived from DAVID.

Term	**P-value (raw)**	**Benjamini-Hochberg FDR**	**Fold Enrichment**
Ubl conjugation	0.0010	0.0048	4.02
Isopeptide bond	0.0016	0.0048	5.02
Phosphoprotein	0.0955	0.1910	1.51

Table 5.2: Statistical Summary of Enriched Functional Terms

Term	**Count**	**Genes**
Ubl conjugation	7	TOP2A, ZEB2, CDCA5, PTCH1, ACTL6A, TAGLN2, HMGCR
Isopeptide bond	6	TOP2A, ZEB2, PTCH1, ACTL6A, TAGLN2, HMGCR
Phosphoprotein	8	TOP2A, ZEB2, CDCA5, PTCH1, ACTL6A, CD248, TAGLN2, HMGCR

Table 5.3: Functional annotation of genes based on enriched terms

Term: Specific functional annotation or pathway.
Count: The number of genes associated with the term.
P-value: The probability of observing the analysis or enrichment by chance alone. A lower value indicates stronger statistical significance.
Benjamini-Hochberg FDR (False Discovery Rate): The adjusted p-value that controls for multiple testing errors. A value less than 0.05 suggests that that the result is unlikely due to random chance.
Fold Enrichment: Indicates how strongly a term is overrepresented in the gene list. A value greater than 1 means the term appears more frequently than expected by chance.

Table 5.2 showcases the relevance of our identified gene set with the selected pathways while Table 5.3 highlights which genes are relevant to which pathway.

Table 5.2 illustrates that ubl conjugation achieved an FDR of 0.0048, which is statistically significant and not likely due to chance. Additionally, fold enrichment validates the claim even to a greater degree by highlighting a solid score of 4.02, and this indicates that the genes related to ubl conjugation are 4 times more frequent in the input gene list. Following this, the Isopeptide bond identified utilizing DAVID further demonstrates strong statistical significance with an FDR of the same 0.0048 but a higher fold enrichment of 5.02. Although the phosphoprotein was markedly enriched by the 8 genes, its FDR is 0.1910, which is borderline. Hence, in this study, it did not receive significant consideration, unlike the other two terms.

Through DAVID integration, our research identified the Ubiquitin-like protein (Ubl) conjugation pathway and the Isopeptide bond. Ubiquitin is a tiny protein present in all eukaryotic cells that serves as a tag or label. Likewise, ubiquitin-like proteins (Ubls) are a family of proteins that function similarly to ubiquitin. The process by which the cell attaches specific tags (Ubl) to other proteins to govern the behavior pattern is called Ubl conjugation. This system is crucial for marking proteins that need to be destroyed, which is led by the proteasome. Furthermore, this intrinsic tagging also involves the transfer of proteins to different parts of the cells or changing the way they work. Due to this specialty, it helps in cell division, DNA repair, and immune response to fight against cancer [14]. Therefore, we may conclude from the enrichment analysis that the identified genes jointly participate in these fundamental regulatory mechanisms.
Besides, some of these well-known ubl proteins are known as SUMO, NEDD8, ISG15, and FAT10 [53]. Studies claimed that the misregulation of the NEDD8 pathway, a specific Ubl conjugation process, increases oncogenic mechanisms, which constitute with cancer [9]. On that account, dysregulation of ubl has been implicated in various diseases, including cancers [10]. Another study shows that PD-1 therapy works better when NEDD8 in TNBC is blocked [70]. So, targeting NEDD8 could be a new way to improve cancer treatment. Hence, we may conclude that ubl conjugation is linked with response to immunotherapy. Consequently, this functional analysis also identified another significant resurgence, which is the isopeptide bond, and it has been identified between the 6 genes from the list. Recent studies also imply that ubl binds to other proteins through the formation of this isopeptide bond [75]. Therefore, the biological significance derived in the context of BC and ICI therapy is profound, suggesting that our gene list collectively contributes to pathways that affect bladder cancer growth.

Integrated Biological Insights and Clinical Implications:

The comprehensive biological analysis of the 11 genes reinforces their relevance to BC and supports their candidacy as potential biomarkers for ICI treatment. Several genes, such as TOP2A, ZEB2, HMGCR, CDCA5, and CD248, are associated with ICI treatment as they have established strong roles in immune modulation, tumor aggressiveness, and checkpoint signaling. For instance, TOP2A is overexpressed in BC and also upregulates PD-L1, which highlights its potential for pembrolizumab therapy. Likewise, ZEB2 indicates a prospective function in resistance to ICI therapy as it is coupled with immune evasion.
Correspondingly, research shows that inhibiting HMGCR with statins can enhance

ICI treatment for many different types of cancers, which denotes its therapeutic relevance in BC. Likewise, the survival response rate facilitated when CD248 and PD-1 considered together, again highlighting its application to BC treatment. Notably, the gene CDCA5 elevates PD-L1 expression and repairs cancerous cell DNA, which reinforces its dual role in oncology. The functional enrichment analysis further revealed significant pathways indicated by these potential biomarkers, which implies that these potential biomarkers collectively shape the tumor microenvironment. For instance, as ubls are designed to control the protein activity, they are linked to the checkpoints such as PD-1/PD-L1/ and CTLA-4. So, misregulating these pathways permits the cancerous cell to escape immune attack. Therefore, considering this pathway further enables clinicians to observe who will benefit from the treatment.

5.2 Analysis and Comparison of Biomarkers with DNNs

This study integrated DNNs to observe their potential in identifying possible biomarkers for BC response to ICI therapy. Although DNNs are well-renowned for their ability to analyse complex datasets, their black-box nature and prone to overfit make them less suitable for high-dimensional datasets with small patient cohorts [55]. Therefore, to justify even further, our study incorporated DNNs to analyze the contrast. The LOOCV value obtained from the DNNs applied is shown in Table 5.4.

Model	**Mean LOOCV Accuracy**
Concrete Autoencoder	0.45
Deep Neural Network with L1 Regularization	0.54

Table 5.4: Mean LOOCV accuracy of different deep learning models

Nevertheless, to strengthen the validity of our framework, we explored other strategic modeling. Here, we incorporated two DNN frameworks: a deep neural network with L1 regularization and a Concrete Autoencoder. Similar to ensemble and individual models, these DNNs were evaluated using LOOCV. While both models select some microRNA but research is still being done to observe the potential of microRNA as biomarkers. Furthermore, we also observed that neither of the model could identify any of the possible biomarkers, which denotes inconsistency in feature selection. Hence, this feature selection strategy implies that DNNs are unable to handle high-dimensional datasets with limited patient cohorts. Additionally, when compared with ensemble-based feature selection methods, DNNs demonstrated lower stability, lower variance, and inadequate biological interpretation, as a result making them less reliable for this kind of research. In essence, DNNs may hold promising results for biomarker identification; however, their limitation is sparse with high-dimensional data and small patient cohorts.

5.3 Interpretation and Explainability of Model Output

In the realm of data science, where AI relies on clarity, our ensemble technique enhances this by explicitly demonstrating how each feature is derived. This is particularly important in clinical fields, where physicians look for strong evidence before administering any treatments. Furthermore, our models' answers are easy to understand and interpret, which makes them highly relevant in a biological context.

Selection of Biologically Meaningful Features

Firstly, through a rigorous ensemble-based feature selection process, this research selects 11 genes which has been identified as potential biomarkers. As a matter of fact, each of the identified genes has been shown in previous studies, which implies our results are not based on random chance; they play key roles in human biology. To elaborate, all of the mentioned genes are linked to oncologic pathways or immune evasion. Besides, a portion of the genes were selected by both ensemble and individual models, which suggests a strong stability. Research shows that genes, such as TOP2A, ZEB2, PTCH1, CDCA5, and CD248, is linked to BC, and these genes play key roles either in immune suppression or tumor progression. Additionally, for some genes, both catergory of intervention have been observed. In brief, our robust techniques identified genes with deep biological significance that can be used for potential pembrolizumab treatment guidance.

Clear Feature Selection Process

A complex architectural system of neural networks often makes it hard to elaborate on how certain results have been achieved. Following that, our simple yet powerful frequency-based majority voting system combined results from four different ensembles and individual models (SVM-RFE, RF-RFE, MI, and LR-RFE). Furthermore, when a gene is picked by the ensemble model, it is considered significant because ensemble models combine the strength of multiple individual models, which results in bias reduction, unlike individual models. Likewise, comparative analysis was conducted with the genes selected by the individual models which further showcase certain genes identified by individual models are also present in ensemble models. Therefore, this transparent integration provides a vivid picture where researchers can see the reasoning for choosing a certain gene, which enables the accurate administration of ICI therapy.

5.4 Discussion

In the era of precision oncology, the core issue is not gathering data but deriving elusive biomarkers for potentially efficacious therapeutic outcomes. This study reinforces the proposal and validation of precise biomarkers that serve as indicators for ICI therapy response. The dataset utilized in this study was acquired from the NCBI's GEO database, which comprises the BC cohort, specifically patients undergoing pembrolizumab-based ICI therapy. The intention behind this study was to demonstrate that ensemble-based feature selection can eliminate or diminish the

effects of the "curse of dimensionality" and accurately identify potential biomarkers while mitigating the limitations of high feature counts and limited sample size. Likewise, the framework provided in this study effectively confronts these issues.
One of the key aspects of this study is that it showcases that the ensemble feature selection-based methods, such as SVM-RFE, RF-RFE, MI, and LR-RFE, perform better than individual models. To make the argument even stronger, it was observed that ensemble models demonstrated better results than deep learning models, such as neural networks with L1 Lasso Regularization and Concrete Autoencoder. This strong comparison with DNNs and singular models signifies that ensemble techniques are well established and provide stable results. Furthermore, our ensemble techniques achieved this stability through aggregation by using both a weighted and unweighted frequency-based voting system. While this mechanism played a crucial role in the robust performance, alongside this, strategic implementation, such as incorporating nested LOOCV and Stratified K-fold CV, was instrumental, contributing further to enhance the generalizability of the finding. Likewise, the nested CV had a significant impact in mitigating the limitations of the dataset, with only 11 samples.
Over and above, through these strategic implementations, 30 genes from each model were taken into account. Subsequently, whilst maintaining the priority of ensemble feature selection high, the literature review-based approach was followed to find the most relevant genes. Additionally, ensemble feature selection was given a priority because features selected by the ensemble are chosen repeatedly across multiple iterations, unlike other techniques, resulting in more stable gene selections. Our analysis also reveals that ensemble feature selection outperforms complex architectures like deep learning methods, particularly in resolving the "curse of dimensionality". Hence, our study underscores the significance of choosing the appropriate modeling technique while considering the dataset rather than solely relying on a model's apparent level of sophistication and algorithmic strength. Instead, our model harnessed the individual strengths and biases of each model using the ensemble approach to select a final feature set for validation that was balanced and well-founded.
One of the key attributes of this study is the biological relevance of the potential biomarkers. Through a robust feature selection strategy, we filtered 11 potential biomarkers, which have prior literature on their effect on immune response and BC. This further solidifies our claims that our ensemble technique is proficient in deriving these important potential biomarkers from a high-dimensional dataset with a small patient cohort. Similarly, this direct link plays a key role in patient stratification and immune response to pembrolizumab. Furthermore, as TOP2A, CDCA5 is heavily linked with PD-L1/PD-1 levels [60] [64], and ZEB2 is also found to be highly correleated to PD-1/ PD-L1 [43], this implies that clinicians should consider TOP2A, CDCA5, and ZEB2 expression levels when administering and during treatment duration of ICI therapy, such as pembrolizumab, for BC patients. Conversely, while the metrics analysed may provide suitable results, this study should be considered exploratory and further investigation is required to solidify our claim.

Chapter 6

Conclusion

6.1 Conclusion

This thesis presents a comprehensive machine learning framework for identifying potential biomarkers in BC patients, with a focus on predicting patient response to pembrolizumab-based ICI therapy. By leveraging both a weighted and unweighted frequency-based majority voting system, which is classified as an ensemble technique, this research identified 11 potential biomarkers. In this study, models such as SVM-RFE, RF-RFE, MI, and LR-RFE have been incorporated in both ensemble and single forms. Additionally, DNNs have been incorporated to assess the performance, which have historically shown to be more precise. Despite this, our study highlights that ensemble techniques are superior when the dataset contains a small patient cohort (N=11) with an exorbitantly high number of features (f=70,523). This study also addressed bias through the cooperation of nested LOOCV and Stratified K-fold CV and weighted and unweighted frequency-based majority voting, which enhanced the models' generality and provided realistic performance estimates. Furthermore, biological validation revealed that there is strong correlation between the identified genes and immune response pathways, suggesting the performance to be much better than what the metrics would suggest.

6.2 Limitations and Future Studies

Although our thesis successfully showed the superiority of ensemble models in circumstances where the patient cohort is small and high-dimensional, further research is needed to assess the model's generalizability in the case of high patient cohorts. In addition, even though the incorporation of nested LOOCV and Stratified K-fold CV reduces bias, it also increases variance, especially with such a small dataset, and this impacted the way the performance metrics had to be observed. This implies that research involving small patient samples and high-dimensional genetic data requires new metrics for assessment, as this type of data is scarce. In conclusion, our thesis addressed the research gap by selecting relevant biomarkers for BC patients in response to ICI therapy.

Bibliography

[1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996, ISSN: 0035-9246. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

[2] J. I. Epstein, M. B. Amin, V. R. Reuter, F. K. Mostofi, B. C. C. Committee, *et al.*, "The world health organization/international society of urological pathology consensus classification of urothelial (transitional cell) neoplasms of the urinary bladder," *The American journal of surgical pathology*, vol. 22, no. 12, pp. 1435–1448, 1998.

[3] R. Maclin and D. Opitz, "Popular ensemble methods: An empirical study," *Journal of Artificial Intelligence Research*, vol. 11, pp. 169–198, 1999. DOI: 10.1613/jair.614.

[4] I. Guyon, J. Weston, S. Barnhill, and et al., "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389–422, 2002. DOI: 10.1023/A:1012487302797.

[5] A. C. Tan and D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification," 2003.

[6] S. A. De Andres, "Variable selection from random forests: Application to gene expression data," *arXiv preprint q-bio/0503025*, 2005. [Online]. Available: https://arxiv.org/abs/q-bio/0503025.

[7] T. Barrett, D. B. Troup, S. E. Wilhite, *et al.*, "Ncbi geo: Mining tens of millions of expression profiles—database and tools update," *Nucleic acids research*, vol. 35, no. suppl_1, pp. D760–D765, 2007.

[8] A. J. Hackstadt and A. M. Hess, "Filtering for increased power for microarray data analysis," *BMC Bioinformatics*, vol. 10, p. 11, Jan. 2009. DOI: 10.1186/1471-2105-10-11.

[9] T. A. Soucy, L. R. Dick, and J. E. Brownell, "The NEDD8 conjugation pathway and its relevance in cancer biology and therapy," *Genes Cancer*, vol. 1, no. 7, Sep. 2010. DOI: 10.1177/1947601910382898.

[10] L. Bedford, J. Lowe, L. R. Dick, R. J. Mayer, and J. E. Brownell, "Ubiquitin-like protein conjugation and the ubiquitin-proteasome system as drug targets," *Nature Reviews Drug Discovery*, vol. 10, no. 1, pp. 29–46, Jan. 2011. DOI: 10.1038/nrd3321.

[11] D. L. Fei, A. Sanchez-Mejias, Z. Wang, *et al.*, "Hedgehog signaling regulates bladder cancer growth and tumorigenicity," *The Journal of Urology*, vol. 188, no. 3, pp. 996–1004, Sep. 2012. DOI: 10.1016/j.juro.2012.07.039.

[12] F. Rapaport, R. Khanin, Y. Liang, and et al., "Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data," *Genome Biology*, vol. 14, no. 9, p. 3158, Sep. 2013. DOI: 10.1186/gb-2013-14-9-r95.

[13] C. Stretch, S. Khan, N. Asgarian, *et al.*, "Effects of sample size on differential gene expression, rank order and prediction accuracy of a gene signature," *PLoS ONE*, vol. 8, no. 6, e65380, 2013. DOI: 10.1371/journal.pone.0065380.

[14] F. C. J. Streich and C. D. Lima, "Structural and functional insights to ubiquitin-like protein conjugation," *Annual Review of Biophysics*, vol. 43, pp. 357–379, 2014. DOI: 10.1146/annurev-biophys-051013-022958.

[15] C.-L. Chen, T. Chung, C.-C. Wu, *et al.*, "Comparative tissue proteomics of microdissected specimens reveals novel candidate biomarkers of bladder cancer," *Molecular Cellular Proteomics*, vol. 14, no. 9, pp. 2466–2478, Sep. 2015. DOI: 10.1074/mcp.M115.051524.

[16] P. Sharma and J. P. Allison, "The future of immune checkpoint therapy," *Science*, vol. 348, no. 6230, pp. 56–61, 2015. DOI: 10.1126/science.aaa8172.

[17] P. Aguiar Jr, L. A. Perry, and G. L. Lopes Jr, *Cost–effectiveness of immune checkpoint inhibitors in nsclc according to pd-l1 expression*, 2016.

[18] S. Tan, C. W. Zhang, and G. F. Gao, "Seeing is believing: Anti-pd-1/pd-l1 monoclonal antibodies in action for checkpoint blockade tumor immunotherapy," *Signal transduction and targeted therapy*, vol. 1, no. 1, pp. 1–4, 2016.

[19] M. S. Thiese, B. Ronna, and U. Ott, "P value interpretations and considerations," *Journal of Thoracic Disease*, vol. 8, no. 9, E928–E931, 2016. DOI: 10.21037/jtd.2016.08.16.

[20] S. Antoni, J. Ferlay, I. Soerjomataram, A. Znaor, A. Jemal, and F. Bray, "Bladder cancer incidence and mortality: A global overview and recent trends," *European urology*, vol. 71, no. 1, pp. 96–108, 2017.

[21] D. T. Silverman, S. Koutros, J. D. Figueroa, L. Prokunina-Olsson, and N. Rothman, "Bladder cancer," in *Schottenfeld and Fraumeni Cancer Epidemiology and Prevention, Fourth Edition*, Oxford University Press, 2017, pp. 977–996.

[22] M. Suryavanshi, J. Sanz-Ortega, D. Sirohi, *et al.*, "S100P as a marker for urothelial histogenesis: A critical review and comparison with novel and traditional urothelial immunohistochemical markers," *Advances in Anatomic Pathology*, vol. 24, no. 3, pp. 151–160, May 2017. DOI: 10.1097/PAP.0000000000000150.

[23] L. T. Kagohara, G. L. Stein-O'Brien, D. Kelley, *et al.*, "Epigenetic regulation of gene expression in cancer: Techniques, resources and analysis," *Briefings in functional genomics*, vol. 17, no. 1, pp. 49–63, 2018.

[24] Q. Li, C. Li, J. Chen, *et al.*, "High expression of long noncoding RNA NORAD indicates a poor prognosis and promotes clinical progression and metastasis in bladder cancer," *Urologic Oncology*, vol. 36, no. 6, 310.e15–310.e22, Jun. 2018. DOI: 10.1016/j.urolonc.2018.02.019.

[25] K. Raza, "Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule," in *U-Healthcare Monitoring Systems*, Elsevier, 2018, pp. 179–196. DOI: 10.1016/B978-0-12-815370-3.00008-6.

[26] N. Sundahl, S. Rottey, D. De Maeseneer, and P. Ost, "Pembrolizumab for the treatment of bladder cancer," *Expert Review of Anticancer Therapy*, vol. 18, no. 2, pp. 107–114, 2018.

[27] F.-H. Wang, X.-J. Ma, D. Xu, and J. Luo, "UPK1B promotes the invasion and metastasis of bladder cancer via regulating the Wnt/-catenin pathway," *European Review for Medical and Pharmacological Sciences*, vol. 22, no. 17, pp. 5471–5480, Sep. 2018. DOI: 10.26355/eurrev_201809_15807.

[28] B. Ghojogh and M. Crowley, "The theory behind overfitting, cross validation, regularization, bagging, and boosting: Tutorial," *arXiv preprint arXiv:1905.12787*, 2019. [Online]. Available: https://arxiv.org/abs/1905.12787.

[29] B. Homet Moreno, M. Dueñas, M. Martínez-Fernández, and J. M. Paramio, *Identification of potential biomarkers of response and resistance to programmed cell death-1 blockade in patients with advanced urothelial tumors*, NCBI Gene Expression Omnibus (GEO), GEO Accession: GSE111636, Available at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111636, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111636.

[30] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLoS ONE*, vol. 14, no. 11, e0224365, 2019. DOI: 10.1371/journal.pone.0224365.

[31] M. Wang and A. Barbu, "Are screening methods useful in feature selection? an empirical study," *PLoS ONE*, vol. 14, no. 9, e0220842, 2019. DOI: 10.1371/journal.pone.0220842.

[32] Z. Wang, A. Lachmann, and A. Ma'ayan, "Mining data and metadata from the gene expression omnibus," *Biophysical reviews*, vol. 11, pp. 103–110, 2019.

[33] S. Zeng, A. Liu, L. Dai, *et al.*, "Prognostic value of TOP2A in bladder urothelial carcinoma and potential molecular mechanisms," *BMC Cancer*, vol. 19, no. 1, p. 604, Jun. 2019. DOI: 10.1186/s12885-019-5814-y.

[34] R. Bai, Z. Lv, D. Xu, and J. Cui, "Predictive biomarkers for cancer immunotherapy with immune checkpoint inhibitors," *Biomarker research*, vol. 8, no. 1, p. 34, 2020.

[35] S. Camelliti, V. Le Noci, F. Bianchi, *et al.*, "Mechanisms of hyperprogressive disease after immune checkpoint inhibitor therapy: What we (don't) know," *Journal of Experimental & Clinical Cancer Research*, vol. 39, pp. 1–20, 2020.

[36] G. Fu, Z. Xu, X. Chen, H. Pan, Y. Wang, and B. Jin, "CDCA5 functions as a tumor promoter in bladder cancer by dysregulating mitochondria-mediated apoptosis, cell cycle regulation and PI3k/AKT/mTOR pathway activation," *Journal of Cancer*, vol. 11, no. 9, pp. 2408–2420, Feb. 2020. DOI: 10.7150/jca.35372.

[37] W. W. Hsieh, "Improving predictions by nonlinear regression models from outlying input data," *arXiv preprint arXiv:2003.07926*, 2020. [Online]. Available: https://arxiv.org/abs/2003.07926.

[38] B. Pes, "Ensemble feature selection for high-dimensional data: A stability analysis across multiple domains," *Neural Computing and Applications*, vol. 32, no. 10, pp. 5951–5973, 2020.

[39] K. Saginala, A. Barsouk, J. S. Aluru, P. Rawla, S. A. Padala, and A. Barsouk, "Epidemiology of bladder cancer," *Medical sciences*, vol. 8, no. 1, p. 15, 2020.

[40] C. P. Wild, E. Weiderpass, and B. W. Stewart, "World cancer report," 2020.

[41] K. C. Arbour, A. T. Luu, J. Luo, *et al.*, "Deep learning to estimate recist in patients with nsclc treated with pd-1 blockade," *Cancer discovery*, vol. 11, no. 1, pp. 59–67, 2021.

[42] V. Buhrmester, D. Münch, and M. Arens, "Analysis of explainers of black box deep neural networks for computer vision: A survey," *Machine Learning and Knowledge Extraction*, vol. 3, no. 4, pp. 966–989, 2021.

[43] B. Dong, J. Liang, D. Li, *et al.*, "Identification of a prognostic signature associated with the homeobox gene family for bladder cancer," *Frontiers in Molecular Biosciences*, vol. 8, p. 688 298, 2021. DOI: 10.3389/fmolb.2021.688298.

[44] J. Hu, C. Cui, W. Yang, *et al.*, "Using deep learning to predict anti-pd-1 response in melanoma and lung cancer patients from histopathology images," *Translational oncology*, vol. 14, no. 1, p. 100 921, 2021.

[45] B. W. Labadie, A. V. Balar, and J. J. Luke, "Immune checkpoint inhibitors for genitourinary cancers: Treatment indications, investigational approaches and biomarkers," *Cancers*, vol. 13, no. 21, p. 5415, 2021.

[46] Y. Li, K. Zhang, F. Yang, *et al.*, "Prognostic value of vascular-expressed PSMA and CD248 in urothelial carcinoma of the bladder," *Frontiers in Oncology*, vol. 11, p. 771 036, Nov. 2021. DOI: 10.3389/fonc.2021.771036.

[47] A. Liu, Z. Zhang, F. Liu, *et al.*, "Identification of a prognostic signature associated with the homeobox gene family for bladder cancer," *Frontiers in Molecular Biosciences*, vol. 8, p. 688 298, Jul. 2021. DOI: 10.3389/fmolb.2021.688298. [Online]. Available: https://www.frontiersin.org/journals/molecular-biosciences/articles/10.3389/fmolb.2021.688298/full.

[48] A. Lopez-Beltran, A. Cimadamore, A. Blanca, *et al.*, "Immune checkpoint inhibitors for the treatment of bladder cancer," *Cancers*, vol. 13, no. 1, p. 131, 2021.

[49] A. Lopez-Beltran, A. Cimadamore, A. Blanca, *et al.*, "Immune checkpoint inhibitors for the treatment of bladder cancer," *Cancers*, vol. 13, no. 1, p. 131, 2021. DOI: 10.3390/cancers13010131.

[50] R. H. Pokhrel, S. Acharya, J.-H. Ahn, *et al.*, "AMPK promotes antitumor immunity by downregulating PD-1 in regulatory T cells via the HMGCR/p38 signaling pathway," *Molecular Cancer*, vol. 20, no. 1, p. 133, Oct. 2021. DOI: 10.1186/s12943-021-01438-x.

[51] L. P. Rhea, S. Mendez-Marti, D. Kim, and J. B. Aragon-Ching, "Role of immunotherapy in bladder cancer," *Cancer Treatment and Research Communications*, vol. 26, p. 100 296, 2021.

[52] S. Borhani, R. Borhani, and A. Kajdacsy-Balla, "Artificial intelligence: A promising frontier in bladder cancer diagnosis and outcome prediction," *Critical reviews in oncology/hematology*, vol. 171, p. 103 601, 2022.

[53] S.-J. Jeon and K.-C. Chung, "Covalent conjugation of ubiquitin-like ISG15 to apoptosis-inducing factor exacerbates toxic stimuli-induced apoptotic cell death," *Journal of Biological Chemistry*, vol. 298, no. 10, p. 102 464, Oct. 2022. DOI: 10.1016/j.jbc.2022.102464.

[54] M. Khalsan, L. R. Machado, E. S. Al-Shamery, *et al.*, "A survey of machine learning approaches applied to gene expression analysis for cancer prediction," *IEEE Access*, vol. 10, pp. 27 522–27 534, 2022.

[55] Y. Wang, H. Chen, X. Jiao, *et al.*, "PTCH1 mutation promotes antitumor immunity and the response to immune checkpoint inhibitors in colorectal cancer patients," *Cancer Immunology, Immunotherapy*, vol. 71, pp. 111–120, 2022. DOI: 10.1007/s00262-021-02966-9.

[56] F. W. Chin, S. C. Chan, and A. Veerakumarasivam, "Homeobox gene expression dysregulation as potential diagnostic and prognostic biomarkers in bladder cancer," *Diagnostics (Basel)*, vol. 13, no. 16, p. 2641, Aug. 2023. DOI: 10.3390/diagnostics13162641.

[57] J. Liu, M. T. Islam, S. Sang, L. Qiu, and L. Xing, "Biology-aware mutation-based deep learning for outcome prediction of cancer immunotherapy with immune checkpoint inhibitors," *NPJ Precision Oncology*, vol. 7, no. 1, p. 117, 2023.

[58] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 2, pp. 757–774, 2023.

[59] Y. Tang, Q. Zhang, Y. Qi, and D. Chen, "PTCH1 mutation as a potential predictor of immune checkpoint inhibitors in gastrointestinal cancer," *Annals of Oncology*, vol. 34, no. Supplement 2, S268, Oct. 2023.

[60] J. Wu, L. Zhang, W. Li, *et al.*, "The role of TOP2A in immunotherapy and vasculogenic mimicry in non-small cell lung cancer and its potential mechanism," *Scientific Reports*, vol. 13, no. 1, p. 10 906, 2023. DOI: 10.1038/s41598-023-38117-6.

[61] H. Eben and M. Bolanle, "Ensemble methods for diabetes prediction: Bagging boosting and stacking," Oct. 2024.

[62] L. Gan, T. Lu, Y. Lu, *et al.*, "Endosialin-positive CAFs promote hepatocellular carcinoma progression by suppressing CD8+ T cell infiltration," *Journal for Immunotherapy of Cancer*, vol. 12, no. 9, e009111, Sep. 2024. DOI: 10.1136/jitc-2024-009111.

[63] Y. He, G. Li, Y. Wu, *et al.*, "Actin like 6a is a prognostic biomarker and associated with immune cell infiltration in cancers," *Discovery Oncology*, vol. 15, no. 1, p. 503, Sep. 2024. DOI: 10.1007/s12672-024-01388-0.

[64] Y. W. Koh, Y. Hwang, S.-K. Lee, J.-H. Han, S. Haam, and H. W. Lee, "The impact of CDCA5 expression on the immune microenvironment and its potential utility as a biomarker for PD-L1/PD-1 inhibitors in lung adenocarcinoma," *Translational Oncology*, vol. 46, p. 102 024, Aug. 2024. DOI: 10.1016/j.tranon.2024.102024.

[65] P. Le, X. Gong, L. Ung, *et al.*, "A robust ensemble feature selection approach to prioritize genes associated with survival outcome in high-dimensional gene expression data," *Frontiers in Systems Biology*, vol. 4, p. 1 355 595, 2024. DOI: 10.3389/fsysb.2024.1355595.

[66] G. Li, C. Li, C. Wang, and Z. Wang, "Suboptimal capability of individual machine learning algorithms in modeling small-scale imbalanced clinical data of local hospital," *PLoS ONE*, vol. 19, no. 2, e0298328, 2024.

[67] G. Li, C. Li, C. Wang, and Z. Wang, "Suboptimal capability of individual machine learning algorithms in modeling small-scale imbalanced clinical data of local hospital," *PLoS ONE*, vol. 19, no. 2, e0298328, 2024. DOI: 10.1371/journal.pone.0298328.

[68] V. W. Lumumba, D. Kiprotich, M. L. Mpaine, N. G. Makena, and M. D. Kavita, "Comparative analysis of cross-validation techniques: LOOCV, K-folds cross-validation, and repeated K-folds cross-validation in machine learning models," *American Journal of Theoretical and Applied Statistics*, vol. 13, no. 5, pp. 127–137, Oct. 2024. DOI: 10.11648/j.ajtas.20241305.13.

[69] V. W. Lumumba, D. Kiprotich, M. L. Mpaine, N. G. Makena, and M. D. Kavita, "Comparative analysis of cross-validation techniques: LOOCV, K-folds cross-validation, and repeated K-folds cross-validation in machine learning models," *American Journal of Theoretical and Applied Statistics*, vol. 13, no. 5, pp. 127–137, 2024. DOI: 10.11648/j.ajtas.20241305.13.

[70] I. Papakyriacou, G. Kutkaite, M. Rúbies Bedós, *et al.*, "Loss of NEDD8 in cancer cells causes vulnerability to immune checkpoint blockade in triple-negative breast cancer," *Nature Communications*, vol. 15, no. 1, p. 3581, Apr. 2024. DOI: 10.1038/s41467-024-47987-x.

[71] R. Theisen, H. Kim, Y. Yang, L. Hodgkinson, and M. W. Mahoney, "When are ensembles really effective?" *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[72] H. Wei, Z. Li, K. Qian, *et al.*, "Unveiling the association between HMG-CoA reductase inhibitors and bladder cancer: A comprehensive analysis using Mendelian randomization, animal models, and transcriptomics," *The Pharmacogenomics Journal*, vol. 24, no. 5, pp. 1–12, Oct. 2024. DOI: 10.1038/s41397-024-00346-x.

[73] A. Younis and J. Gribben, "Immune checkpoint inhibitors: Fundamental mechanisms, current status and future directions," *Immuno*, vol. 4, no. 3, pp. 186–210, 2024.

[74] M. T. Marrone, J. E. Reuss, A. Crawford, *et al.*, "Statin use with immune checkpoint inhibitors and survival in nonsmall cell lung cancer," *Clinical Lung Cancer*, vol. 26, no. 3, pp. 201–209, May 2025. DOI: 10.1016/j.cllc.2024.12.008.

[75] R. Mousa, D. Shkolnik, Y. Alalouf, and A. Brik, "Chemical approaches to explore ubiquitin-like proteins," *RSC Chemical Biology*, vol. 6, pp. 492–509, Feb. 2025. DOI: 10.1039/D4CB00220B.

[76] National Center for Biotechnology Information (NCBI), *PTCH1 patched 1 [Homo sapiens (human)]*, https://www.ncbi.nlm.nih.gov/gene/5727, Gene ID: 5727. Accessed on 18 June 2025., 2025.