A Novel Machine Learning Framework for Identifying Predictive Biomarkers of FGFR Targeted Therapy in Breast Cancer

Rishiraj Sinharay - 32435851

Monash University

FIT5128

Supervisor: A/Prof. Lan K. Nguyen

Co-Supervisor: Dr. Sungyoung Shin

CONTENTS

PART I: GENERAL LITERATURE REVIEW

1. INTRODUCTION

Breast cancer is one of the most common cancers among women. Global Cancer Statistics (GLOBOCAN) data shows breast cancer is diagnosed in 1 out of 4 (24.2%) women worldwide [3]. It is a complex disease characterised by an interplay between lifestyle, environmental, and genetic factors and has continually changing treatment recommendations. From diagnosis, breast cancer is considered a systemic disease (affects the entire body) and requires different treatment approaches such as surgery, radiation, chemotherapy, and/or targeted therapy [2,4]. Radiation therapy uses high-energy radiation, and chemotherapy uses drugs to attack cancer cells but, unfortunately, affects other cells in the body in the process, which leads to numerous side effects. Targeted therapy is a newer form of therapy where the drug only attacks the cancerous cells in the body, making it more effective and less toxic than radiation and chemotherapy. Inhibitors against several pathways that promote cell signalling for rapid cell growth have shown promising results in clinical trials [7]. Targeted therapy is based on previous genetic analysis and complements conventional diagnostic methods [7]. Extensive studies on breast cancer and recent advances in ‘omics’ have helped in the identification of numerous molecular targets and the development of novel therapeutics [7,17].

Fibroblast growth factor receptors (FGFR) mutations are a key candidate that correlates with a high level of drug sensitivity in breast cancer [8]. Their role in cancer cell proliferation, survival, and angiogenesis has made them a promising target for breast cancer therapy. These receptors bind to the FGF family of proteins. Gene mutations in FGFRs are common in certain subtypes of breast cancer, which cause increased FGFR signalling, resulting in the rapid growth of cancer cells. In FGFR-targeted therapy, FGFR inhibitors aim to attack the FGFR signalling pathways and inhibit these signals. The inhibitors bind to the FGF receptors and block the signalling activity, thereby inhibiting the growth of cancer cells.

Breast cancer treatment depends on multiple variables; thus, single-gene biomarkers are insufficient to make accurate decisions. Drug targets alone are poor therapeutic indicators, which makes identifying reliable biomarkers a challenging task, not only for cytotoxic drugs but for targeted therapy as well [11,12]. Hence, a reliable multi-gene biomarker panel is preferred to make accurate decisions [6,11,12]. Biomarkers help evaluate the benefits of an inhibitor in targeted therapy. They also help evaluate a drug's toxic side effects in chemotherapy [9]. Correct selection of biomarkers helps assess the malignancy and disease remission level, predict the response to therapy, and monitor how the therapy advances [4][9].

The increase in experimental methods for biosample profiling and the collection of clinical and health record data over time offer promising opportunities for biomarker discovery [15]. With the advent of precision oncology, we can utilise patients’ genomic makeup and data from previous genetic analyses for therapeutic decisions [11, 12]. Based on this data, numerous statistical models and machine learning approaches have been used to build models that can predict resistance and response to drugs both in clinical and preclinical settings [12]. The fundamental phases in the computational approach to biomarker and drug response prediction are (a) drug response quantification, (b) feature selection and dimensionality

reduction, (c) machine learning model fitting, and (d) model evaluation [12].

This project will integrate protein expression data. sourced from the Cancer Cell Line Encyclopedia (CCLE) and AZD4547 FGFR inhibitor drug data collected from the DepMap portal [34]. These data are utilised to build and train novel machine-learning models for deriving biomarker panels that are sensitive to FGFR inhibitors.

2. SUBSTANTIVE LITERATURE REVIEW

2.1. Breast cancer overview

Breast cancer occurs in the breast tissue when cells in the human breast mutate and generate uncontrollably, creating a mass of tissue (tumour) [1]. Breast cancer is classified into different subtypes based on the presence or absence of specific hormone receptors and/or genetic mutations, such as Human Epidermal Growth Factor Receptor 2 (HER2) amplification. Diagnosis of breast cancer typically involves a combination of clinical evaluation and imaging techniques, such as mammogram, breast ultrasound, breast MRI, CT scan, PET scan [13,14,15,16], along with biopsy, which confirms the presence of cancer cells or tumours in the breast tissue. Biomarkers are generally classified into two subgroups: prognostic and predictive. Prognostic biomarkers predict patient clinical outcomes irrespective of the treatment, whereas predictive biomarkers predict the response of a patient to a particular therapeutic intervention and are associated with the sensitivity of the cancer tumour to the type of therapy [6]. Breast cancer is one of the most malignant tumours in women, with studies being carried out for over ten years to identify its predictive biomarkers [9].

Estrogen receptor (ER(**α**)), Progesterone receptor (PgR), and HER2 are well-established biomarkers in breast cancer and are crucial in predicting prognosis [4,5,6]. ER(**α**) expression is the most important biomarker as it provides the sensitivity index to endocrine treatment as ER is a direct target for endocrine therapies. ER status also predicts response to chemotherapy in a neoadjuvant setting. PgR expression is strongly dependent on the presence of ER. Tumors expressing PgR but not ER are very rare (< 1%) and are also made to undergo endocrine therapy. Overexpression of HER2 protein can be detected in about 15% of all primary breast cancers [6,13]. HER2 factor has mixed prognostic and predictive significance and is the target of the monoclonal antibody trastuzumab. Its amplification gives a good predictive response to anti-HER2 therapy [6]. Ki-67, Cyclin D1, and Cyclin E are some of the emerging biomarkers in breast cancer and are all associated with the cell cycle [10]. However, due to the heterogeneity of breast cancer and the vast number of variables that influence treatment response and outcome, single-gene biomarkers are insufficient for the decision-making process [5,11]. Multi-gene biomarker panels are more likely to capture the complex tumor-drug response [11]. Many multigene signatures have been identified that aim to outperform traditional biomarkers. These help in defining specific characteristics and the possibility for individual treatment optimisation [6]. Oncotype DX and MammaPrint are examples of multigene tests that are commercially available for breast cancer [4,6]. OncotypeDx is a predictive biomarker and is based on a 21-gene panel to predict the risk of recurrence in patients being treated with chemotherapy for early-stage ER-positive breast cancer [6]. MammaPrint is a prognostic biomarker that measures the expression levels of 70 genes to classify the tumour for treatment decisions [6].

The assessment of the tumour at the time of diagnosis, i.e. staging, helps decide the most suitable treatment option [1,4]. Therapeutic options for patients with breast cancer have been mostly based on histological properties. In addition to this, extensive research on linking the correlation between molecular

characteristics of subtypes of breast cancer and their therapeutic outcomes has shown the importance of molecular and genomic heterogeneity in patients receiving the same treatment [18]. Radiation therapy and chemotherapy are based on a one-size-fits-all approach. In radiation therapy, high-energy radiation is used to attack the DNA inside the cancer cells, which kills the cells. Even though it is highly effective, radiation exposure can lead to the development of cancer in other parts of the body and may cause skin irritation and other long-term side effects. Chemotherapy uses drugs to target cells in the human body which are dividing rapidly. These include cancer cells but also healthy cells like blood cells, hair follicles, reproductive cells, and cells in the digestive tract. This causes hair loss, nausea and vomiting, fatigue, and other side effects. With the development of precision oncology, treatment decisions consider patients' genomic makeup and tumor-site agnostic molecular aberration biomarkers [12]. Validation of these biomarkers has led to the identification of therapeutic targets. Targeted therapy is a newer approach in which agents/ inhibitors are designed to attack and block proteins or molecules specific to the cancer cells (targets), which makes them less toxic compared to chemotherapy. ER and HER2 are therapeutic targets for which a lot of research has been done to develop drugs for the treatment of breast cancer [19].

2.2. FGFR signalling in breast cancer

The Fibroblast Growth Factor Receptors (FGFRs) are one of the actionable targets for targeted therapy of breast cancer. The FGFR signalling system controls various fundamental biological processes, including tissue formation, angiogenesis, and tissue regeneration. It is an evolutionarily conserved signalling cascade [20,21,22]. Numerous physiological cellular functions, including embryonic development, differentiation, proliferation, survival, migration, and angiogenesis, are regulated by FGF and FGFR signalling [20]. The FGFR family is made up of four TKIs (FGFR1-4), each of which has an extracellular, transmembrane, and cytoplasmic domain [24]. Regulation of FGF signalling is necessary to ensure balanced receptor stimulation. The FGFR pathway has been altered in various ways in cancer, including (i) gene amplification or post-transcriptional regulation that results in receptor overexpression; (ii) FGFR mutations that produce receptors that are either constitutively active or exhibit a reduced dependence on ligand binding for activation [29]; (iii) translocations that result in the expression of FGFR-fusion proteins with constitutive FGFR kinase activity; (iv) alternative splicing [23,24]. Breast cancer growth and progression have been linked to abnormal FGFR signalling, which is typically accompanied by elevated or altered expression of FGF ligands and genetic altercations. For example, in between 8% and 15% of all breast cancers and between 16% and 27% of luminal type B breast cancers, FGFR1 is amplified. These amplifications are connected to FGFR1 overexpression, endocrine treatment resistance, and poor prognosis [25]. Triple-negative breast cancer frequently has FGFR2 amplifications [29], and FGFR inhibitors are known to have a powerful effect on these abnormalities [5]. Studies have also linked a higher incidence of sporadic post-menopausal breast cancer to FGFR2. Breast cancer has also been linked to the fusions of the FGFR1-3 gene, frequently involving numerous gene partners. The FGFR3 and FGFR4 amplifications in breast cancer are, however, extremely uncommon. Furthermore, through retaining tumor-initiating cells, FGFR2 has been shown to play a critical role in enhancing breast cancer tumorigenicity [23,26,27,29]. Therefore, targeting FGFR signalling has become a promising therapeutic approach in the treatment of breast cancer.

In clinical studies, many FGFR inhibitors, including AZD4547, BGJ398, and dovitinib,

have been investigated for the treatment of breast cancer. AstraZeneca's AZD4547 is a very effective and selective FGFR1-3 inhibitor. During a phase I trial, 5 of 20 patients with tumours harbouring FGFR signalling abnormalities showed minimal activity [20]. Patients with a high level of FGFR amplification had better efficacy. One of the eight patients with FGFR1-amplified breast cancer in a phase II multicenter proof-of-concept study evaluating AZD4547 responded to the inhibitor [27]. Although AZD4547 did not fulfil the primary endpoint, preliminary activity signals appeared to be limited to tumours with FGFR-activating mutations and fusions. Different somatic FGFR mutations may confer varying amounts of signalling potency and/or oncogene dependency [30,31]. Similar to this, in the phase I investigation, only one patient with FGFR1-amplified breast cancer displayed tumour regression when receiving NVP-BGJ398 [27]. Dovitinib had antitumor activity in breast cancer cell lines with FGFR amplification [25]. It also significantly raised plasma levels of FGF23, indicating FGFR1 inhibition [21,22,25,27]. In FGFR1-amplified breast tumours versus nonamplified breast cancers, dovitinib has stronger anticancer effects [25].

2.3. Current challenges in FGFR-targeted therapy

The clinical success of FGFR-targeted therapeutics has been constrained despite encouraging preclinical results because of a number of variables like drug resistance [27, 29], toxicity, and lack of patient stratification. The challenges now centre on choosing patients who are most likely to benefit from these treatments, enhancing the efficacy of therapies through the development of novel potent compounds and combinational techniques, and overcoming FGFR inhibitor-related toxicities [20]. This is due to the fact that FGFR signalling is extremely complicated and interacts with other signalling pathways, which can counteract FGFR inhibition and encourage resistance [35].

The development of effective targeted therapy in breast cancer has been hindered by the lack of reliable biomarkers that can predict response to treatment. To increase the efficacy and decrease the toxicity of FGFR-targeted medicines, it is crucial to develop trustworthy biomarkers that can predict response. FGFR inhibitors can be more effectively prescribed to individuals through reliable and personalised biomarkers. Biomarkers are also used to track the effectiveness of treatment and spot the development of drug resistance.

Due to the complexity of FGFR signalling and the absence of well-validated biomarkers, it is difficult to develop predictive biomarkers for FGFR-targeted therapy. Large-scale clinical trials that are well-planned and have uniform procedures for sample collection, processing, and analysis are necessary for the validation of biomarkers. These trials are costly, time-consuming, and might require collaboration between multiple organisations. In order to create reliable predictive biomarkers for FGFR-targeted therapy, a large investment in clinical research infrastructure and resources would be needed. The heterogeneity of breast cancer, which can result in differences in FGFR expression and activation across various tumor subtypes and stages, presents another challenge. As a result, it might be difficult to find a single biomarker that can reliably predict responsiveness to FGFR inhibitors. Combining multiple biomarkers or creating a composite biomarker signature that integrates the clinical and molecular characteristics of the tumour would help in increasing the precision of response prediction [11]. Multiple parameter combinations have resulted from (i) high-throughput molecular profiling followed by bioinformatics analysis and/or (ii) the sensible combination of established markers with others of significant biological value [5,6]. Careful analysis of biomarker data from

clinical trials may improve the capacity to identify the best patient population for FGFR-targeted treatments [22].

2.4. Machine Learning Frameworks for Response Biomarker Discovery

Large-scale genomic, transcriptomic, and proteomic data sets from breast cancer patients treated with FGFR inhibitors may be analysed by machine learning frameworks to produce such response biomarkers. These frameworks can recognise molecular signatures linked to therapy response or resistance and use those signatures to create tailored treatment plans for specific patients [11, 12]. The accuracy and reproducibility of ML frameworks are dependent on the quality and amount of the available data, but they have the potential to identify molecular signatures linked to response or resistance to FGFR inhibitors. The clinical trial, preclinical, and other data can be used to build ML algorithms that can forecast treatment outcomes and pinpoint the most predictive biomarkers [12]. The challenge of high data dimensionality, juxtaposed with a limited number of samples, often referred to as the "curse of dimensionality" [11] or the p>>n problem [16], is a pervasive issue in pharmacogenomics data. This problem arises primarily due to the exponential increase in molecular features as compared to the finite and often limited number of available biological samples. Too many features create a space with a relatively high dimension, and the data points will be sparsely distributed, leading to statistically unstable models and thus causing overfitting of data [16]. Machine learning algorithms require feature selection strategies to handle this. To solve the difficulty of drug response prediction, a move from developing different predictive models for each drug based on separate data subsets to developing a single model based on all available data has been proposed as a more effective strategy. Within this centralised model, a framework is employed in which all medications share certain factors while others are adapted to the unique qualities of each drug. This strategy, as explained by Adam et al. (2020) [12], leverages both pharmacological similarities and differences to improve the accuracy and efficiency of drug response forecasts. Using a unified modelling framework allows for a more comprehensive knowledge of pharmacological behaviours and provides a simplified technique for continuously investigating drug response prediction.

The data type, the research topic, and the goal of the ML framework will all influence the decision of which ML algorithm to use for predictive biomarker discovery. While unsupervised algorithms are utilised for clustering and dimensionality reduction, supervised learning algorithms are frequently used for classification and prediction tasks. Decision trees, random forests, support vector machines (SVM), and neural networks are some of the frequently used methods for biomarker discovery. In particular, decision trees can manage missing data and are simple to read, but they may overfit the training set of data. High-dimensional data can be handled by random forests, although they can be computationally expensive. SVMs can be effective and efficient in handling complex data, but careful parameter tuning may be necessary. Neural networks are capable of picking up complicated patterns but may also overfit and require a large amount of training data.

ML frameworks can objectively and methodically analyse huge and complex datasets, which is one of its main advantages. They help in finding complex patterns and relationships in data and can also find biomarkers that are most indicative of a patient’s reaction to treatment. Additionally, ML can be used to create models that forecast how each patient will respond to a certain course of treatment, enabling the development of personalised oncology.

ML has emerged as a strong method for identifying predictive biomarkers in cancer therapy in recent years. For example, Shin et al. (2023) created a novel computational framework on MATLAB that combines supervised machine learning-based biomarker discovery with Boolean algebra-based signature derivation in order to identify a predictive multi-gene biomarker signature of HSP90-targeted treatment for prostate cancer [11]. There have been considerable advances in ML-based response biomarker discovery in the context of FGFR-targeted therapy for breast cancer. These employ a variety of ML frameworks, including Boolean and Bayesian models, to find potential biomarkers that can predict responsiveness to FGFR-targeted therapy. The OncotypeDx and MAMMAPRINT are gene expression-based breast cancer biomarkers. Both tests look at the expression of certain genes in breast tumour tissues to predict cancer recurrence and the efficacy of chemotherapy. MAMMAPRINT measures the expression of 70 genes and generates a risk of recurrence score. OncotypeDX measures the expression of 21 genes and generates a recurrence score [4, 12]. These tests help in making educated decisions about the best treatment decisions for patients with early-stage, hormone receptor-positive breast cancer.

3. SUMMARY OF STATE OF THE ART

Breast cancer is the most common malignant tumour among women worldwide, which necessitates a variety of therapeutic modalities, including surgery, radiation, chemotherapy, and targeted therapy. Targeted therapy is a newer form of treatment that targets cancerous cells, making it more efficient and less toxic than radiation and chemotherapy. TKIs (Tyrosine Kinase Inhibitors) are an example of inhibitors that are directed against several targets, including FGFR, which are important candidates and correlate with a high level of drug sensitivity in breast cancer. FGFR inhibitors attack the FGFR signalling pathways and block the signals which stop the proliferation of cancer cells. Individual biomarkers are insufficient to make informed decisions regarding the treatment of breast cancer because of its heterogeneity; therefore, a multi-gene biomarker panel is preferred. Biomarkers assist in assessing the advantages of an inhibitor drug in targeted therapy, measuring the degree of malignancy and disease remission, forecasting the therapeutic response, and tracking the progress of the therapy. The development of experimental techniques for biosample profiling as well as the collection of clinical and health record data over time, present exciting potential for the discovery of new biomarkers. Precision oncology has made it possible to use patient genomic data from previous genetic analyses to guide treatment choices. Numerous statistical models and machine learning techniques have been applied to this data to develop models that can forecast medication resistance and response in both clinical and preclinical contexts. There is still a need for a trustworthy multi-gene biomarker panel to make reliable decisions for breast cancer treatment for targeted therapy, even though biomarkers have been extensively studied in breast cancer, and recent advances in ‘omics’ have helped identify numerous molecular targets and develop novel therapeutics.

4. PLAN FOR RESEARCH PROJECT

4.1. Aim and Tasks

Targeted therapies have shown encouraging results in treating breast cancer patients in recent years, but the development of novel therapeutics using multi-gene biomarker panels is critical for making reliable decisions. The primary goal is to analyse protein expression and drug response data using novel machine learning frameworks to derive response biomarkers that can predict the response of FGFR inhibitors. As such, the aims of the research are:

1. Using ML frameworks to derive predictive biomarkers for FGFR-targeted therapy in breast cancer.
2. Building an ML model pipeline and testing the derived biomarkers on an independent cell line dataset of breast cancer patients.

To achieve these, we can divide the project into the following tasks:

a. Exploratory data analysis on the primary datasets, like data cleaning and wrangling to prepare the final dataset.
b. Utilise machine learning algorithms to determine which features best predict the AUC value (feature selection).
c. Use qualitative knowledge regarding biomarkers and their relationship to cell lines corresponding to breast cancer to filter out accurate biomarkers.
d. Build and train ML models and test the selected panel of biomarkers against independent cell line datasets of patients with breast cancer.

4.2. Datasets and Computational Requirement

The datasets used in this project are the protein expression data which is obtained from the Cancer Cell Line Encyclopedia (CCLE) [32], which has 12755 rows of data and 16384 feature columns that give rise to the p>>n problem. Drug response data for the AZD4547 FGFR inhibitor was gathered from the DepMap portal [34]. By combining the data on protein expression with drug response data for FGFR inhibitors, we will build the final dataset and use it to build and train novel machine-learning models to identify biomarker panels that are responsive to FGFR inhibitors.

The exploratory data analysis, model fitting, feature selection, and classification will be carried out using Python version 3.9 with the help of packages like numpy, pandas, and machine learning frameworks like sci-kit-learn, etc.

4.3. Ethical Considerations

The CCLE and GDSC datasets used in this research project are publicly available datasets and have been anonymised by the publishers. They have also been used in many other research papers. Thus, there are limited concerns regarding privacy and ethics.

5. CONCLUSION

The identification of multi-gene response biomarker panels among breast cancer patients has the potential to improve treatment outcomes and aid in better decision-making in breast cancer. Furthermore, the application of machine learning frameworks in biomarker identification has the potential to improve drug development and reduce the time and cost associated with traditional therapeutic approaches. Hence, this research project aims to use novel ML frameworks to analyse drug response and protein expression data in breast cancer patients and identify biomarkers that can be used to predict the response of FGFR inhibitors. The initiative is expected to contribute to the development of targeted medicines for breast cancer patients by identifying accurate multi-gene biomarker panels. Machine learning algorithms and qualitative knowledge about biomarkers and their relationship to breast cancer cell lines will be used to develop a ML pipeline and test the identified biomarkers against an independent breast cancer cell line dataset.

6. REFERENCES

[1] Cleveland Clinic. (2022, January 21). *Breast Cancer: Causes, Stage, Diagnosis & Treatment*. Cleveland Clinic. https://my.clevelandclinic.org/health/diseases/3986-breast-cancer

[2] Ely, S., & Vioral, A. N. (2007). Breast Cancer Overview. *Plastic Surgical Nursing*, *27*(3), 128–133.

https://doi.org/10.1097/01.psn.0000290281.48197.ae

[3] Perou, C. M., Sørlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., ... & Botstein, D. (2000). Molecular portraits of human breast tumours. *nature*, *406*(6797), 747-752.

[4] *Breast Cancer Biomarkers | ARUP Consult*. (n.d.). Arupconsult.com. https://arupconsult.com/content/breast-cancer

[5] Patani, N., Martin, L.-A., & Dowsett, M. (2013). Biomarkers for the clinical management of breast cancer: International perspective. *International Journal of Cancer*, *133*(1), 1–13. https://doi.org/10.1002/ijc.27997

[6] Weigel, M. T., & Dowsett, M. (2010). Current and emerging biomarkers in breast cancer: prognosis and prediction. *Endocrine-Related Cancer*, *17*(4), R245–R262. https://doi.org/10.1677/erc-10-0136

[7] Mohamed, A., Krajewski, K., Cakar, B., & Ma, C. X. (2013). Targeted Therapy for Breast Cancer. *The American Journal of Pathology*, *183*(4), 1096–1112. https://doi.org/10.1016/j.ajpath.2013.07.005

[8] Higgins, M. J., & Baselga, J. (2011). Targeted therapies for breast cancer. *Journal of Clinical Investigation*, *121*(10), 3797–3803. https://doi.org/10.1172/jci57152

[9] Nalejska, E., Mączyńska, E., & Lewandowska, M. A. (2014). Prognostic and Predictive Biomarkers: Tools in Personalized Oncology. *Molecular Diagnosis & Therapy*, *18*(3), 273–284. https://doi.org/10.1007/s40291-013-0077-9

[10] Ross, J. S., Linette, G. P., Stec, J., Clark, E., Ayers, M., Leschly, N., Symmans, W. F., Hortobagyi, G. N., & Pusztai, L. (2003). Breast cancer biomarkers and molecular medicine. *Expert Review of Molecular Diagnostics*, *3*(5), 573–585. https://doi.org/10.1586/14737159.3.5.573

[11] Shin, S.-Y., Centenera, M. M., Hodgson, J. T., Nguyen, E. V., Butler, L. M., Daly, R. J., & Nguyen, L. K. (2023). A Boolean-based machine learning framework identifies predictive biomarkers of HSP90-targeted therapy response in prostate cancer. *Frontiers in Molecular Biosciences*, *10*. https://doi.org/10.3389/fmolb.2023.1094321

[12] Adam, G., Rampášek, L., Safikhani, Z., Smirnov, P., Haibe-Kains, B., & Goldenberg, A. (2020). Machine learning approaches to drug response prediction: challenges and recent progress. *Npj Precision Oncology*, *4*(1). https://doi.org/10.1038/s41698-020-0122-1

[13] mayo clinic. (2019). *Breast cancer - Diagnosis and treatment - Mayo Clinic*. Mayoclinic.org. https://www.mayoclinic.org/diseases-conditions/breast-cancer/diagnosis-treatment/drc-20352475

[14] *Tests to diagnose breast cancer | Cancer Research UK*. (2017). Cancerresearchuk.org. https://www.cancerresearchuk.org/about-cancer/breast-cancer/getting-diagnosed/tests-diagnose

[15] Cancer.net. (2019, June 11). *Breast Cancer - Diagnosis*. Cancer.net. https://www.cancer.net/cancer-types/breast-cancer/diagnosis

[16] Diaz-Uriarte, R., Gómez de Lope, E., Giugno, R., Fröhlich, H., Nazarov, P. V., Nepomuceno-Chamorro, I. A., Rauschenberger, A., & Glaab, E. (2022). Ten quick tips for biomarker discovery and validation analyses using machine learning. *PLOS Computational Biology*, *18*(8), e1010357. https://doi.org/10.1371/journal.pcbi.1010357

[17] Waks, A. G., & Winer, E. P. (2019). Breast Cancer Treatment. *JAMA*, *321*(3), 316. https://doi.org/10.1001/jama.2018.20751

[18] Jhan, J.-R., & Andrechek, E. R. (2017). Triple-negative breast cancer and the potential for targeted therapy. *Pharmacogenomics*, *18*(17), 1595–1609. https://doi.org/10.2217/pgs-2017-0117

[19] Mohamed, A., Krajewski, K., Cakar, B., & Ma, C. X. (2013). Targeted Therapy for Breast Cancer. *The American Journal of Pathology*, *183*(4), 1096–1112. https://doi.org/10.1016/j.ajpath.2013.07.005

[20] Touat, M., Ileana, E., Postel-Vinay, S., André, F., & Soria, J.-C. (2015). Targeting FGFR Signaling in Cancer. *Clinical Cancer Research*, *21*(12), 2684–2694. https://doi.org/10.1158/1078-0432.CCR-14-2329

[21] Ye, T., Wei, X., Yin, T., Xia, Y., Li, D., Shao, B., Song, X., He, S., Luo, M., Gao, X., He, Z., Luo, C., Xiong, Y., Wang, N., Zeng, J., Zhao, L., Shen, G., Xie, Y., Yu, L., & Wei, Y. (2014). Inhibition of FGFR signaling by PD173074 improves antitumor immunity and impairs breast cancer metastasis. *Breast Cancer Research and Treatment*, *143*(3), 435–446. https://doi.org/10.1007/s10549-013-2829-y

[22] André, F., & Cortés, J. (2015). Rationale for targeting fibroblast growth factor receptor signaling in breast cancer. *Breast Cancer Research and Treatment*, *150*(1), 1–8. https://doi.org/10.1007/s10549-015-3301-y

[23] Dienstmann, R., Rodon, J., Prat, A., Perez-Garcia, J., Adamo, B., Felip, E., Cortes, J., Iafrate, A. J., Nuciforo, P., & Tabernero, J. (2014). Genomic aberrations in the FGFR pathway: opportunities for targeted therapies in solid tumors. *Annals of Oncology*, *25*(3), 552–563. https://doi.org/10.1093/annonc/mdt419

[24] Sobhani, N., Ianza, A., D'Angelo, A., Roviello, G., Giudici, F., Bortul, M., Zanconati, F., Bottin, C., & Generali, D. (2018). Current Status of Fibroblast Growth Factor Receptor-Targeted Therapies in Breast Cancer. *Cells*, *7*(7), 76. https://doi.org/10.3390/cells7070076

[25] André, F., Bachelot, T., Campone, M., Dalenc, F., Perez-Garcia, J. M., Hurvitz, S. A., Turner, N., Rugo, H., Smith, J. W., Deudon, S., Shi, M., Zhang, Y., Kay, A., Porta, D. G., Yovine, A., & Baselga, J. (2013). Targeting FGFR with dovitinib (TKI258): preclinical and clinical data in breast cancer. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *19*(13), 3693–3702. https://doi.org/10.1158/1078-0432.CCR-13-0190

[26] Xie, Y., Su, N., Yang, J., Tan, Q., Huang, S., Jin, M., Ni, Z., Zhang, B., Zhang, D., Luo, F., Chen, H., Sun, X., Feng, J. Q., Qi, H., & Chen, L. (2020). FGF/FGFR signaling in health and disease. *Signal Transduction and Targeted Therapy*, *5*(1). https://doi.org/10.1038/s41392-020-00222-7

[27] Babina, I. S., & Turner, N. C. (2017). Advances and challenges in targeting FGFR signalling in cancer. *Nature Reviews Cancer*, *17*(5), 318–332. https://doi.org/10.1038/nrc.2017.8

[28] Higgins, M. J., & Baselga, J. (2011). Targeted therapies for breast cancer. *Journal of Clinical Investigation*, *121*(10), 3797–3803. https://doi.org/10.1172/jci57152

[29] Jhan, J.-R., & Andrechek, E. R. (2017). Triple-negative breast cancer and the potential for targeted therapy. *Pharmacogenomics*, *18*(17), 1595–1609. https://doi.org/10.2217/pgs-2017-0117

[30] Chae, Y. K., Hong, F., Vaklavas, C., Cheng, H. H., Hammerman, P., Mitchell, E. P., Zwiebel, J. A., Ivy, S. P., Gray, R. J., Li, S., McShane, L. M., Rubinstein, L. V., Patton, D., Williams, P. M., Hamilton, S. R., Mansfield, A., Conley, B. A., Arteaga, C. L., Harris, L. N., & O'Dwyer, P. J. (2020). Phase II Study of

AZD4547 in Patients With Tumors Harboring Aberrations in the FGFR Pathway: Results From the NCI-MATCH Trial (EAY131) Subprotocol W. *Journal of Clinical Oncology*, *38*(21), 2407–2417. https://doi.org/10.1200/jco.19.02630

[31] De Luca, A., Frezzetti, D., Gallo, M., & Normanno, N. (2017). FGFR-targeted therapeutics for the treatment of breast cancer. *Expert Opinion on Investigational Drugs*, *26*(3), 303–311. https://doi.org/10.1080/13543784.2017.1287173

[32] *Gygi Lab @ HMS*. (n.d.). Gygi.hms.harvard.edu. Retrieved April 24, 2023, from https://gygi.hms.harvard.edu/publications/ccle.html

[33] *Search results for : fgfr - Cancerrxgene - Genomics of Drug Sensitivity in Cancer*. (n.d.). Www.cancerrxgene.org. Retrieved April 24, 2023, from https://www.cancerrxgene.org/search?query=FGFR

[34] *DepMap: The Cancer Dependency Map Project at Broad Institute*. (n.d.). Depmap.org. https://depmap.org/portal/

[35] Mossahebi-Mohammadi, M., Quan, M., Zhang, J.-S., & Li, X. (2020). FGF Signaling Pathway: A Key Regulator of Stem Cell Pluripotency. *Frontiers in Cell and Developmental Biology*, *8*. https://doi.org/10.3389/fcell.2020.00079

PART II: THE RESEARCH PAPER

ABSTRACT

Breast cancer, a global health problem and a leading malignancy among women is a disease in which cells in the breast grow out of control due to a combination of various hereditary, lifestyle, and environmental factors. Traditional treatments like radiation and chemotherapy have effectively fought the disease but have significant side effects. Targeted therapy presents a promising path by attacking cancer cells and causing minimum side effects. Fibroblast Growth Factor Receptors (FGFR) are one of the key targets in the targeted therapy for Breast Cancer but lack significant biomarkers. The heterogeneity of breast cancer makes it difficult to rely on a single biomarker for making therapeutic decisions. The following study seeks to develop a novel machine-learning framework to evaluate proteomics data combined with drug response data on FGFR inhibitors and derive a robust multi-gene biomarker set capable of forecasting how FGFR inhibitors will react, thus advancing tailored cancer care.

Keywords: Machine learning, breast cancer, targeted therapy, feature selection, predictive biomarker, FGFR inhibitor, precision oncology, bootstrapping

1. INTRODUCTION

Breast cancer is a significant global health challenge due to its high prevalence among women [10]. According to the Global Cancer Statistics (GLOBOCAN) findings, breast cancer affects around 25% of women worldwide [1], highlighting its extensive global prevalence. Despite the progress made in diagnostic and therapeutic methods, which have resulted in enhanced survival rates, the complex characteristics of the disease necessitate a continuous development of treatment approaches due to the interplay of genetic, environmental, and lifestyle variables. Traditionally, the primary approaches for treating breast cancer have consisted of surgery, radiation therapy, and chemotherapy [2,3]. Nevertheless, although essential, these systemic approaches often lead to adverse side effects detracting from the patient's well-being.

In recent years, targeted therapy has gained prominence due to its focus on specific molecular pathways, offering a promising outlook [4,5]. Within the spectrum of targeted therapies, inhibitors that specifically target the Fibroblast Growth Factor Receptors (FGFR) have exhibited significant potential [6]. The FGFR protein, due to its significant role in cancer cell proliferation, survival, and angiogenesis, has established itself as a desirable target for therapeutic interventions [6]. Gene mutations in FGFRs, specifically prevalent in certain breast cancer subtypes, increase FGFR signalling activity, facilitating the accelerated proliferation of cancer cells [28]. The appeal of FGFR-targeted therapy stems from its high degree of precision, as the inhibitors selectively attach to the FGF receptors, blocking their signalling function and thereby curbing cancer cell growth.

Nevertheless, using these inhibitors in a clinical setting faces several obstacles, largely due to the wide range of genetic variations in breast cancer. The heterogeneity of breast cancer further emphasises that singular biomarkers are inadequate for treatment decisions [7]. In contrast, there is a preference for utilising a multi-gene biomarker panel as it offers a comprehensive depiction of the illness condition and the possible therapeutic responses [7,8,9]. Biomarkers provide valuable insights into the efficacy of targeted therapy inhibitor drugs and play a crucial role

in assessing drug toxicity during chemotherapy [10]. The selection of appropriate biomarkers can yield useful insights into the severity of a disease, treatment response, and progression [3,10].

The current era of technical progress and the emerging discipline of precision oncology has presented an unprecedented prospect for using patients' genetic data to make educated therapeutic decisions [8,9,11]. The integration of machine learning into this framework has the potential to enhance the optimisation process, providing robust models for the prediction of drug resistance and response [9]. This study is focused on the convergence of protein expression data obtained from the Cancer Cell Line Encyclopedia (CCLE) and AZD4547 drug response dataset (PRISM Repurposing [84]) derived from The Cancer Dependency Map Project at Broad Institute (DepMap portal) [31] to combine both. This synthesis aims to employ innovative machine-learning techniques to identify biomarker panels responsive to FGFR inhibitors.

2. BACKGROUND

2.1. Breast Cancer, Current Treatments, and Limitations

Breast Cancer is a complex and heterogeneous disease characterised by the unregulated proliferation of mutated cells in the breast tissue, resulting in the development of a tumour [12]. The heterogeneity of the disease is evident via the presence of several subtypes that are categorised according to specific hormone receptors and genetic alterations [7].

The diagnostic methods for breast cancer are extensive, combining clinical evaluation and a range of imaging techniques such as mammography, breast ultrasonography, magnetic resonance imaging (MRI), computed tomography (CT) scans, and positron emission tomography (PET) scans [11,13,14,15]. Biopsies serve as a conclusive means to verify the presence of cancer cells or tumours within the breast tissue. Biomarkers are crucial in managing breast cancer and are commonly categorised into prognostic and predictive groups. Prognostic biomarkers can predict clinical outcomes regardless of the therapy administered, whereas predictive biomarkers play a crucial role in assessing a patient's response to specific treatments [7].

Current treatments for breast cancer exhibit a wide variety of diversity and complexity since they are customised to address the unique characteristics and progression of the illness in individual patients [17]. Conventional therapeutic approaches include surgery for tumour removal, radiation therapy using high-energy radiation to eradicate cancer cells, and chemotherapy, where drugs target rapidly dividing cells, including cancer cells [2,3]. Nevertheless, it is important to note that these traditional therapies have the potential to impact healthy cells, hence resulting in systemic side effects. The advancement of precision oncology has introduced a new era of targeted therapy that aims to target cancer cells selectively, therefore minimising the side effects on the entire system [21,39]. The therapeutic interventions are guided by the genetic profile of the patient and the identification of specific molecular aberrations as biomarkers, resulting in the development of individualised treatment strategies [9].

However, these advanced treatments also have their shortcomings. One of the primary obstacles involves the identification and verification of reliable biomarkers. While existing biomarkers play a key role in guiding treatment decisions, they frequently fail to fully capture the complex dynamics that impact a tumor's response to therapy [8,16]. The reliance on single-gene biomarkers is insufficient, highlighting the pressing need for developing panels that incorporate multiple genes to enhance the accuracy and reliability

of forecasts about treatment responses and prognosis [16, 7].

2.2. FGFR Inhibitors in Breast Cancer

FGFR inhibitors are becoming increasingly prominent in managing breast cancer, mostly owing to their targeted approach [16], which sets them apart from conventional treatments such as chemotherapy and radiation. FGFRs are crucial to many cellular processes, including cell proliferation, survival, and differentiation [18,19,20]. Aberrant FGFR signalling has been implicated in advancing cancer, making it a potential target for target therapy [22,23,24,25,26].

FGFR inhibitors, such as AZD4547 [26], BGJ398, and dovitinib [29], have been formulated to selectively target and suppress the FGFR signalling system, which in turn restricts the excessive cell growth commonly observed in cancer [18]. This alternative option offers a more personalised approach. It lessens toxicity for patients [27,28], aligning with the principles of precision oncology as the therapies administered are specifically suited to each patient's unique genetic and molecular characteristics.

Patient stratification is a significant problem in the context of FGFR inhibitors [22,26]. The efficacy of FGFR inhibitors can exhibit significant variability across individuals due to the heterogeneity of breast cancer. Identifying individuals most likely to benefit from FGFR inhibitor-based targeted therapy [18] requires dependable biomarkers and predictive models. Due to the disease's inherent complexity, developing predictive biomarkers presents a significant challenge in clinical settings. The absence of thoroughly validated biomarkers and the intricate nature of FGFR signalling pathways hamper the accuracy of identifying suitable patients. A thorough examination of biomarker data derived from clinical trials can enhance the ability to identify the best patient population for FGFR-targeted therapies [20].

2.3. Machine Learning Frameworks for Predictive Biomarker Discovery

In the field of response biomarker development, machine learning (ML) frameworks have had a substantial impact since the accumulation of high-throughput omics data began, especially for patients undergoing targeted therapy like FGFR inhibitors [8,9]. They play a crucial role in using extensive genomic, transcriptomic, and proteomic information to discover molecular patterns linked to the effectiveness or resistance of therapeutic interventions [9]. Current ML-based approaches focus on enhancing model performance and prediction accuracy rather than identifying experimentally testable biomarkers. In this study, we develop a novel ML pipeline that identifies predictive biomarkers. The critical factors of data quality and volume significantly influence the accuracy and repeatability of these frameworks. The issue of high data dimensionality, sometimes referred to as the "curse of dimensionality" [8] or the p>>n problem [15], is a persistent and recurring obstacle. Within this particular context, the extensive range of molecular features significantly outnumbers the quantity of accessible biological samples. As a result, this creates data scarcity and overfitting challenges, undermining the model's stability and dependability [15].

To address the issue, the use of feature selection techniques plays a crucial role in the field of machine learning, particularly in the context of biomarker discovery. In addition, Instead of employing individual prediction models for each treatment based on restricted data, it is more efficient to utilise a unified model that has been trained on a pan-cancer dataset. This model has certain shared characteristics across all drugs while also possessing distinct attributes specific to each drug [9]. Machine learning methods, including decision trees, random forests, support vector

machines (SVM), and neural networks, have demonstrated proficiency in handling high-dimensional data [8]. However, it is crucial to execute meticulous tuning to mitigate overfitting and improve the performance of the models. Each method possesses distinct advantages and limitations.

The ability of ML frameworks to analyse and understand extensive information is evidence of their adaptability and precision. This ability allows them to uncover nuanced patterns and associations that are crucial for personalised oncology. Examples such as OncotypeDx and MAMMAPRINT highlight the advancement and effectiveness of ML in customising treatment strategies, evaluating the likelihood of cancer recurrence, and appraising the efficacy of chemotherapy [3,7,9]. These ML applications provide a foundation for therapeutic choices based on insights from data analysis. These advancements exemplify the intersection of technology and medicine, where the analytical capabilities of ML intersect with the complex, individualised requirements of cancer patients, offering tailored and efficient treatment plans.

2.4. Plan for Research

The research plan is around utilising ML frameworks to optimise the identification of predictive biomarkers for FGFR-targeted therapy in breast cancer. The intricate and heterogeneous nature of breast cancer has rendered singular biomarkers inadequate for making well-informed treatment decisions. Hence, the primary objective is to utilise the data to develop a novel ML framework to identify predictive biomarkers for FGFR-targeted therapy in breast cancer. The general workflow of the research is shown in Figure 1.

We start with a comprehensive analysis of pan-cancer cell line data collected from CCLE [30] and AZD4547 FGFR [32] inhibitor drug dataset (PRISM Repurposing [84]) collected from the DepMap portal using exploratory data analysis (EDA) techniques. The initial phase involves data cleaning and data wrangling to build a pan-cancer dataset containing names of cell lines as rows and protein names as columns.

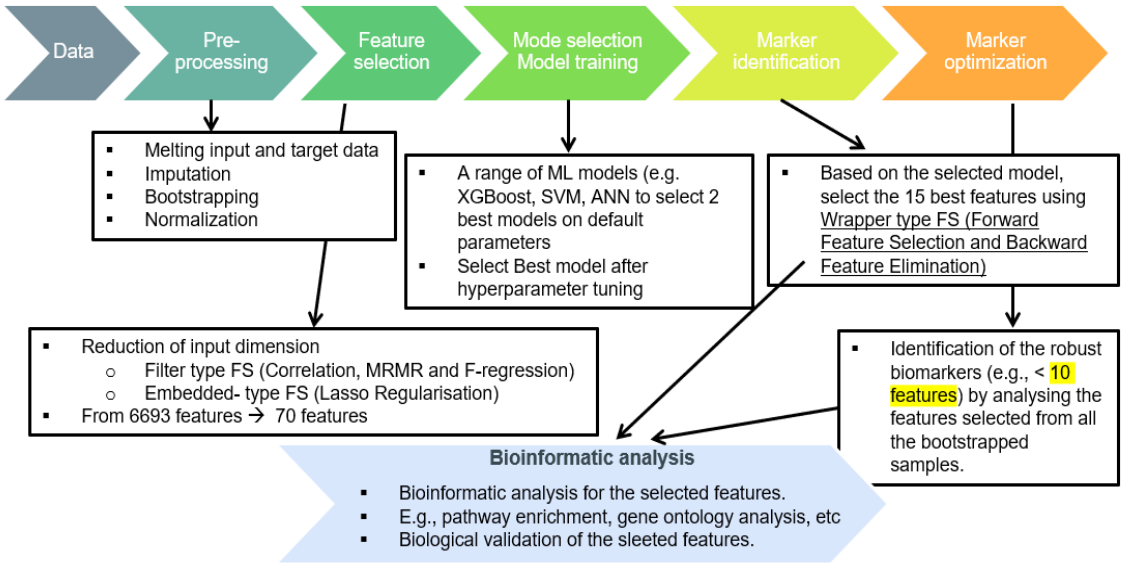


Figure 1. Overview of Predictive Biomarker Identification Pipeline

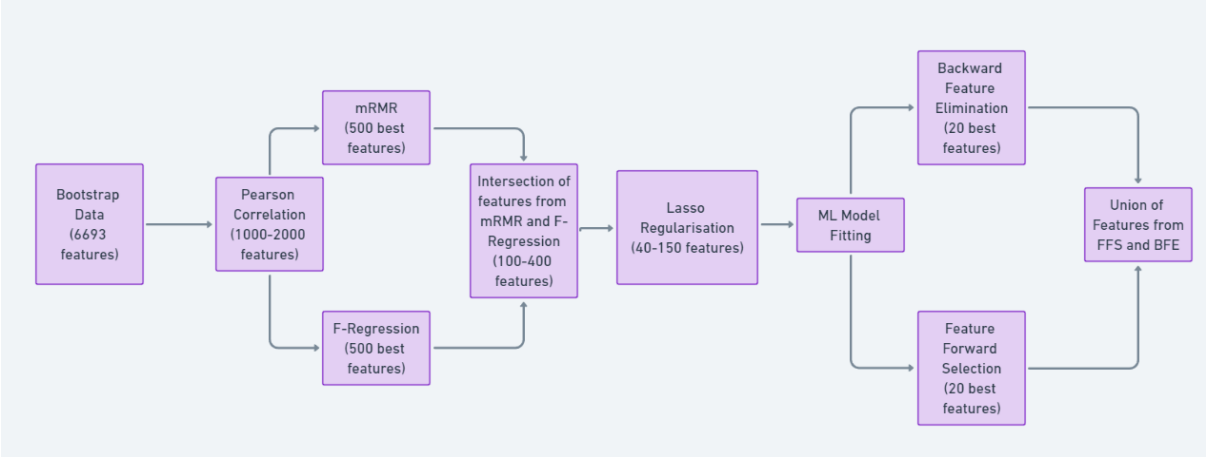


Figure 2. Feature Selection Pipeline for Filtering Relevant Proteins

This is followed by the feature selection phases to choose the most relevant features or variables from the dataset used for machine learning modelling, often ranging from 40 to 150. Then, we further refine the selected features to identify the best features (genes) by applying machine learning techniques. The chosen features undergo bioinformatic analysis, which provides comprehensive biological insights into potential predictive biomarkers. The feature selection pipeline is shown in Figure 2.

The study will look at certain biological indicators and how they connect to cancer cell types to increase how effectively this method works. By doing so, we will ensure that the indicators chosen are statistically significant and essential in a real-world clinical situation. The strategy will be designed to address issues such as too much complicated data or the possibility of drawing particular conclusions from it.

The outcome of this research is expected to be a comprehensive biomarker panel consisting of many genes. To ensure the robustness of the genes selected, we carry out multiple iterations of feature selection and model fitting by bootstrapping the wrangled dataset. This gives us numerous biomarker panels, which can be analysed for the final multi-gene panel of biomarkers for breast cancer. Enrichr, a web-based gene set enrichment analysis application, will be utilised to perform thorough research using the last multi-gene panel [33,34,35]. This strategy aims to ascertain the biological processes inherently associated with our chosen genes. This study will offer valuable insights and significantly improve our comprehensive understanding of the biological functions of the selected biomarker genes.

3. METHODOLOGY

3.1. Data Sources, Collection, and Preprocessing

The cancer cell line data were meticulously obtained from the Cancer Cell Line Encyclopedia (CCLE) database, and drug response data for the AZD4547 FGFR inhibitor was gathered from the DepMap portal (PRISM Repurposing [84]). Both datasets were integrated to create a complete pan-cancer dataset, where each row represents a particular cancer cell line, while the columns count the proteins and corresponding drug dosage.

The missing values were imputed using a Gibbs sampler-based left-censored missing value imputation approach to enhance the completeness and reliability of the dataset

[43,44,45]. We utilised a drug response data dataset, specifically the Area Under Curve (AUC) value, for AZD4547 as target data for machine learning. This represents the overall effect of the drug.

Before the implementation of feature selection techniques, the consolidated pan-cancer dataset had a total of 6692 columns (*input data for machine learning*), each denoting a distinct protein, with an additional column for the AUC values, which is the target column. The dataset consists of 316 cancer cell lines, with each cell line represented as an independent row. This dataset showcases the interaction of different proteins with the AZD4547 FGFR inhibitor in various cancer cell lines.

To ensure the robustness of the selected features using ML frameworks, we bootstrapped the above dataset around 1000 times and got as many different multi-gene biomarker panels as possible. We will later analyse these biomarker panels and select the most robust biomarkers. For each bootstrapped sample, The data is standardised to a Z-score with a mean of 0 and a standard deviation of 1.

3.2. Feature Selection based on Filter and Embedded Methods

In the initial phase of feature selection, we adopt a filter-type method for selecting features. The Pearson correlation analysis is a popular option among filter-type feature selection methods [85]. It identifies the linear relationship between input data (for example, protein expression) and target data (such as the drug response, AUC).

We assumed that the features with a significant correlation between expression and drug response play a more important role in predicting drug response. Therefore, we chose the features with a p-value of less than 0.05 in the correlation analysis. Note that eliminating unnecessary or irrelevant features is essential for enhancing efficiency in subsequent stages of machine learning.

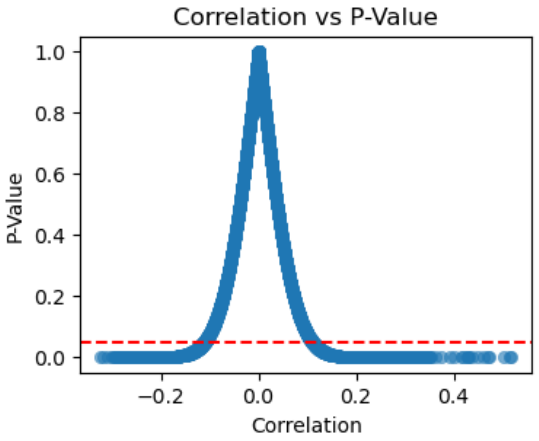


Figure 3. Correlation v/s p-value plot for a particular iteration where each dot represents a feature. The dotted line represents p=0.05.

The features selected from the Pearson correlation are then subjected to further stages of feature selection using the MRMR (Maximum Relevance and Minimum Redundancy) and F-regression algorithms. Consequently, we chose the top 500 features from each of the two algorithms. The MRMR algorithm aids in providing a balanced assessment of each feature's significance relative to the target variable while also measuring the level of redundancy among the features [35,36]. It accomplishes this by simultaneously maximising the mutual information between each feature and the target variable and the depreciation of the information shared among the features. In contrast, F-regression assesses the individual linear relationships between each variable and the target using an F-statistic [86]. Only features displaying statistically significant linear relationships are retained. Subsequently, we combined mRMR and F-regression by extracting shared feature sets.

Next, we applied the Lasso regularisation algorithm to refine the selected features further. The hyperparameters of regularisation, such as the lambda value (the alpha value in Python), are optimised through a thorough

cross-validation process [40]. In other words, we evaluate various alpha values to determine which maximises the model's performance score. As a result, we identified less than 100 features that are used for model training in each bootstrap iteration. The feature selection technique employed in this study is comprehensive and intricate, aiming to derive a collection of statistically significant features and carefully refined features to enhance the model's effectiveness and explanatory capacity.

3.3. Machine Learning: Model Selection, Training, and Validation

For the machine learning and evaluation of the model performance, we employed three different models: XGBoost (Extreme Gradient Boosting) [41], MLPRegressor (Multi-Layer Perceptron) [40], and SVR (Support Vector Regression) [40]. Initially, these models were implemented with their default parameters to establish a baseline for performance evaluation. Subsequently, we found that two models demonstrated superior performance compared to the rest. The two leading models underwent further intensive hyperparameter tunning, utilising RandomizedSearchCV (a Python library), which was enhanced by KFold cross-validation [40].

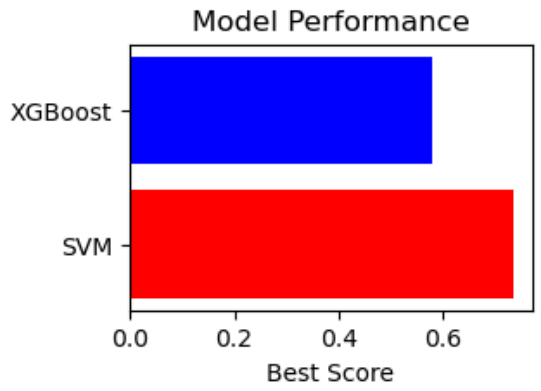


Figure 4. Model performance of a particular bootstrap iteration

A criterion used to assess the model's performance is the R-squared score, which quantifies the correlation between the predicted value and the target values. The model that attained the highest R-squared score was deemed to be the best. This score displays its exceptional ability to explain the variability of the target variable.

The outcome of this rigorous model selection and optimisation process is the identification of ML models that are not only statistically robust but also finely calibrated, which offer superior performance and nuanced interpretability tailored to the dataset's specific characteristics and intricacies.

3.4. Identification of Biomarkers Based on Wrapper Feature Selection Methods

To further refine the selected features from the previous step, we employed the machine learning-based feature selection method, called the wrapper-type method. We selected one of the best-performing models whose hyperparameters are tuned and optimised against the dataset.

For the implementation, the Sequential Feature Selector method of the Sci-kit learn library was utilised, and Feature Forward Selection (FFS) and Backward Feature Elimination (BFE) were carried out [40].

FFS starts with zero initial features and gradually includes those that significantly boost the model’s performance until an established threshold is reached. Conversely, the BFE method begins with a complete feature set and systematically removes the less significant features to retain the most important ones, contributing to correct predictions.

As a result, a total of 20 prominent features or proteins were identified and preserved during each iteration of both the FFS and BFE processes. These procedures were customised to suit the characteristics of each distinct bootstrapped dataset. In this stage, we build a

capable model by incorporating key features and ensuring that it is based on reliable statistical and real-world data. At the end of each bootstrap iteration, we kept the features collected from FFS and BFE in a data frame. This allows us to evaluate the results thoroughly.

3.5. Bioinformatic Analysis of the Selected Features

The features isolated from the wrapper methods were combined and organised. We executed a series of informatics analyses to gain a more comprehensive biological understanding of them. The features/proteins and frequencies are summarised in Figure 5.

We utilised Enrichr, a comprehensive web-based tool for gene set enrichment analysis, to conduct a gene enrichment analysis on this resulting set of genes (which correspond to the proteins). Specifically, this analysis facilitates the identification of the most significantly enriched (over-represented) gene sets among a list of given genes. These gene sets could represent biological processes, pathways, molecular functions, diseases, or any type of relevant gene set [33,34,35].

In the initial phase of the analysis, the focus was to analyse genes acquired by filter methods of feature selection, such as Pearson correlation, mRMR, and F-regression within a specified bootstrap sample. Following that, the genes obtained using embedded feature selection (Lasso Regularisation) were tested in the same bootstrap sample. This two-step strategy allows for a thorough examination, utilising both filter and embedded feature selection strategies.

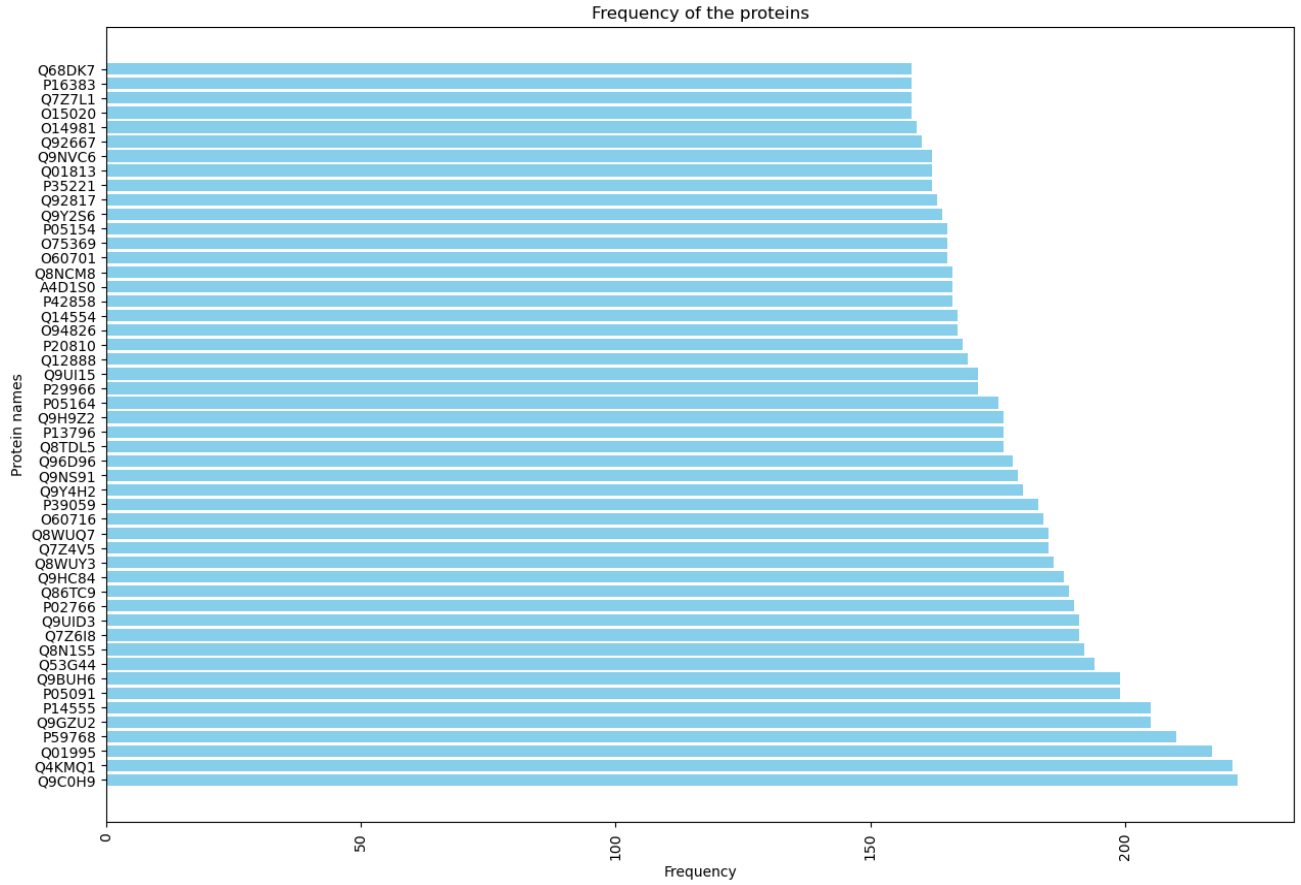


Figure 5. The top 50 proteins and their frequencies selected after bootstrapping

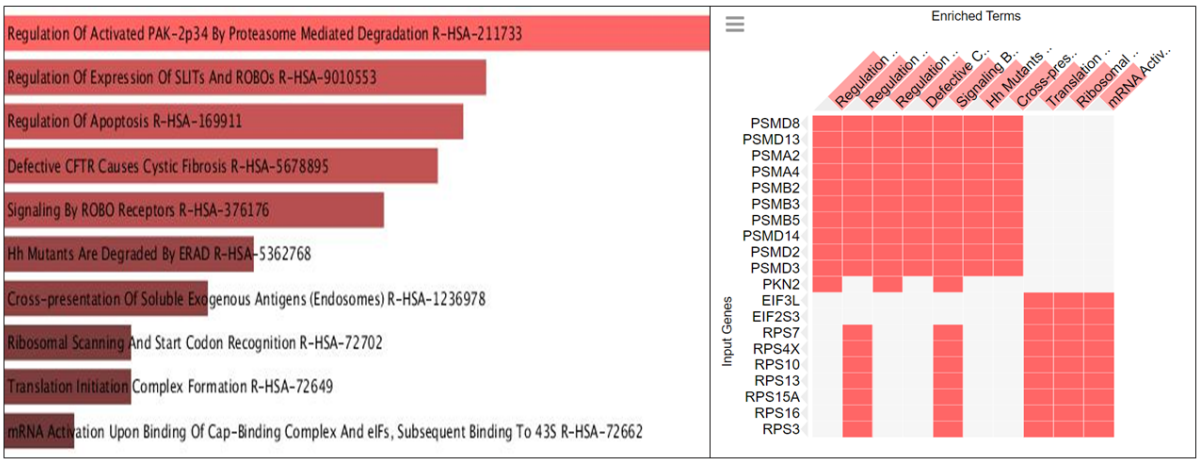
363 distinct proteins and their related genes were found after over 1000 bootstrap iterations. The Enrichr tool was used for gene enrichment analysis to explore deeper into the relevance of these genes. Several gene subsets were investigated, ranging from 363 to 5 genes. The subgroups were chosen based on how frequently each gene occurred in the bootstrap rounds, emphasising the resilience and relevance of each gene. These frequency categories provided as a standard for assessing the biological significance of the genes, allowing for a more nuanced understanding of their functions and effects within the larger natural context. Enrichr was used in this broad technique to extensively examine the complicated linkages and pathways connected with these genes, creating a solid foundation for future enquiries and analysis.

4. RESULTS AND DISCUSSION

The gene sets derived from each stage of the feature selection pipeline were meticulously examined using Enrichr. For this, the gene sets obtained from the filter-type and embedded-type feature selection methods were separately subjected to enriched pathway analysis, providing a more nuanced understanding of the biological pathways they were involved in. Then, the gene set derived from combining the final gene set at the end of all the bootstrap iterations was analysed through different stages to better understand the results.

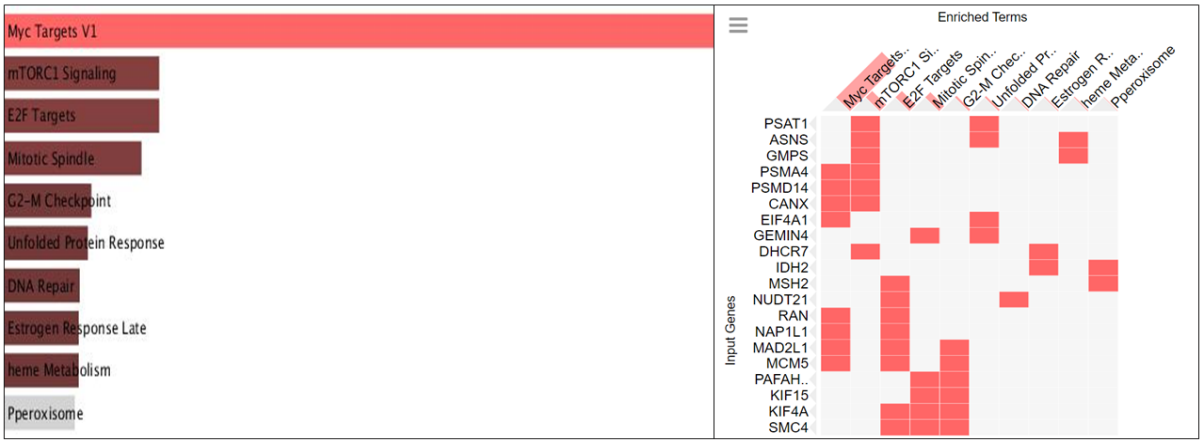
4.1. Enrichment Analysis of Genes Obtained from Filter and Embedded Feature Selection Methods

A total of 247 genes were obtained by applying filter feature selection methods in a designated bootstrap iteration. The pathways associated with the gene sets are shown in Figure 6 and Figure 7. Notably, the Regulation of Expression of SLITs and ROBOs, seen to be enriched in Figure 6(a), has promise in the context of targeted treatment for breast cancer. The reduction of CXCR4 expression effectively hinders the migration and invasion of cancer cells. Activating the Slit2/Robo1 pathway has potential as a viable therapeutic strategy [46]. Furthermore, it has been observed to influence advanced cancer stages by interacting with proteins such as E-cadherin and β-catenin. This presents an additional avenue for potential therapeutic interventions in the context of breast cancer [46]. The gene sets associated with this route contribute to its enrichment and are observed in other enriched pathways, as Figure 6(b) depicts. The Myc Targets V1 pathway, which is highly enriched, as seen in Figure 7, has been found to play a role in breast cancer, explicitly concerning cancer aggressiveness, immune response, and patient survival outcomes [47,48]. The MYC gene particularly exhibits a strong association with recurrent modifications in triple-negative breast cancer (TNBC). Therefore, exploring the therapeutic potential of targeting MYC has emerged as a feasible strategy for treating triple-negative breast cancer (TNBC) [49]. Both the datasets have a few genes in common for the enriched pathways.



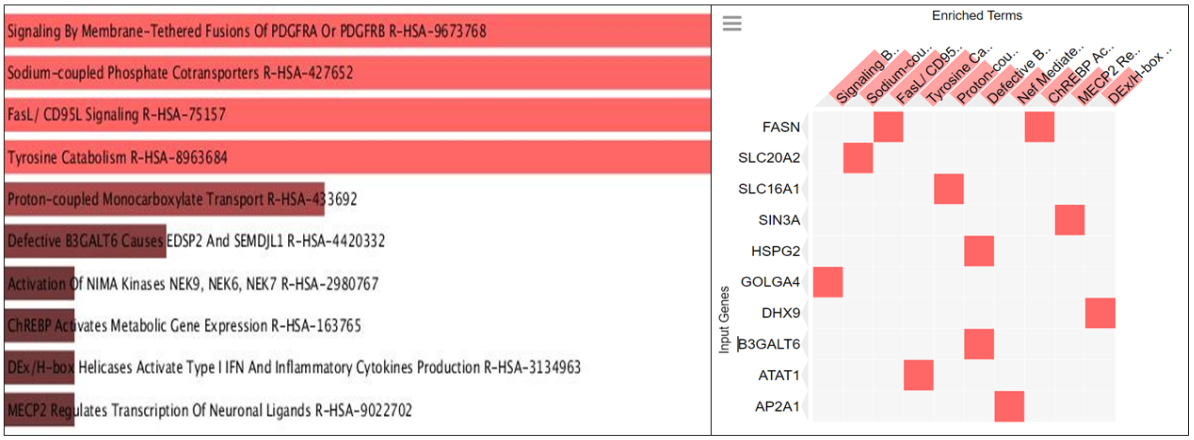
(a) (b)

Figure 6. Pathways enriched in Reactome database by genes obtained from Filter FS methods (a) Ranking of various pathways. (b) Grid analysis of genes present in the pathways.



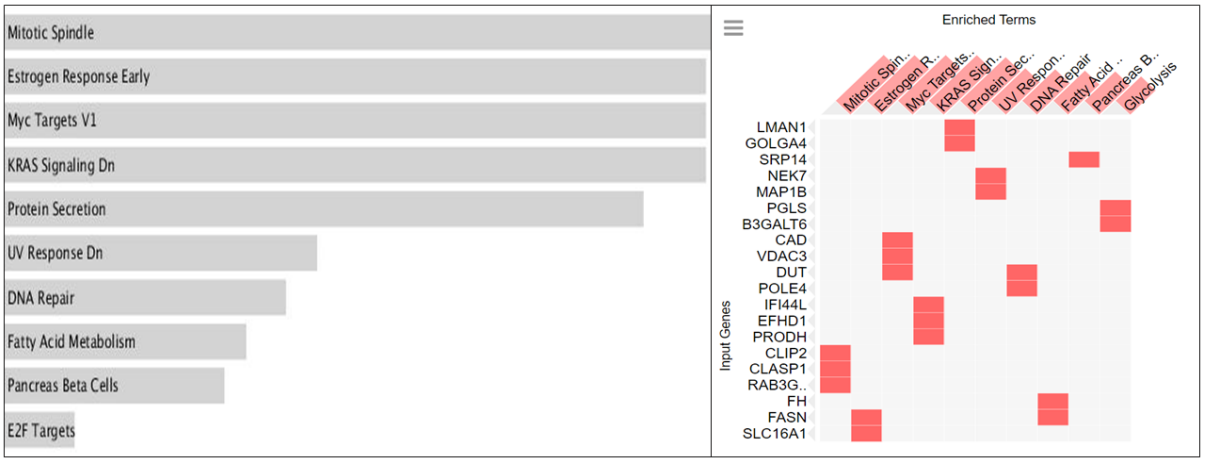
(a) (b)

Figure 7. Pathways enriched in Hallmark database by genes obtained from Filter FS methods (a) Ranking of various pathways. (b) Grid analysis of genes associated with the pathways



(a) (b)

Figure 8. Pathways enriched in Reactome database by genes obtained from Embedded FS (a) Ranking of various enriched pathways. (b) Grid analysis of genes associated with the pathways.



(a) (b)

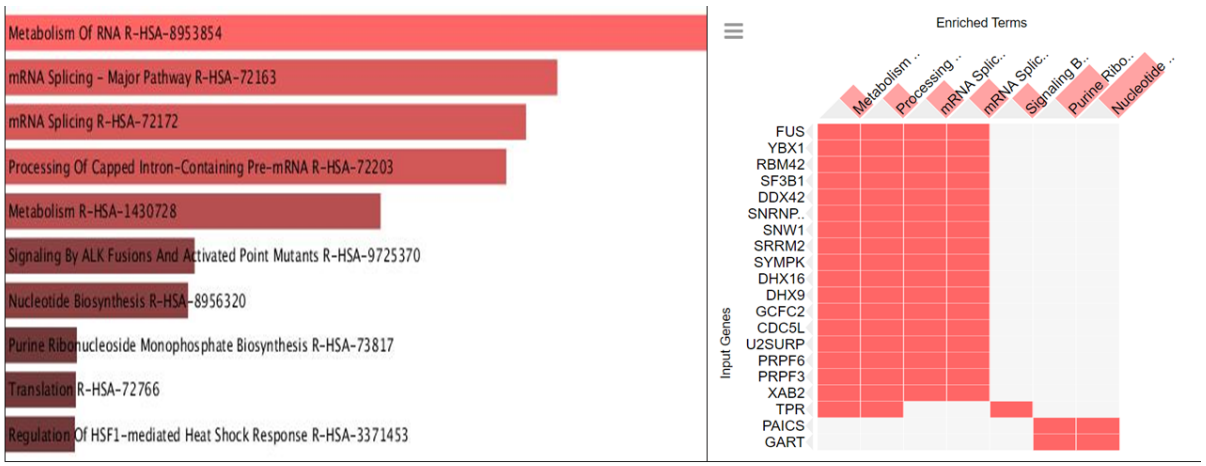
Figure 9. Pathways enriched in Hallmark database by genes obtained from Embedded FS (a) Ranking of various pathways. (b) Grid analysis of genes associated with the pathways.

89 genes collected from embedded feature selection, utilising Lasso regularisation are shown in Figure 8 and Figure 9.

The Reactome database shows that Signalling By Membrane-Tethered Fusions Of PDGFRA Or PDGFRB is highly enriched for the specified gene set (Figure 8). This pathway is known to be associated with several cancer pathologies, including breast cancer. The connection between PDGFRA and PDGFRB and their involvement in oncogenesis and treatment resistance establishes them as significant targets in cancer therapies [50]. There is some enrichment noted in the "Estrogen Response Early" pathways within the Hallmark database (Figure 9), which is a route that has been linked to breast cancer [51,52]. As discussed before, the "Myc Targets V1" pathway is linked to breast cancer [49].

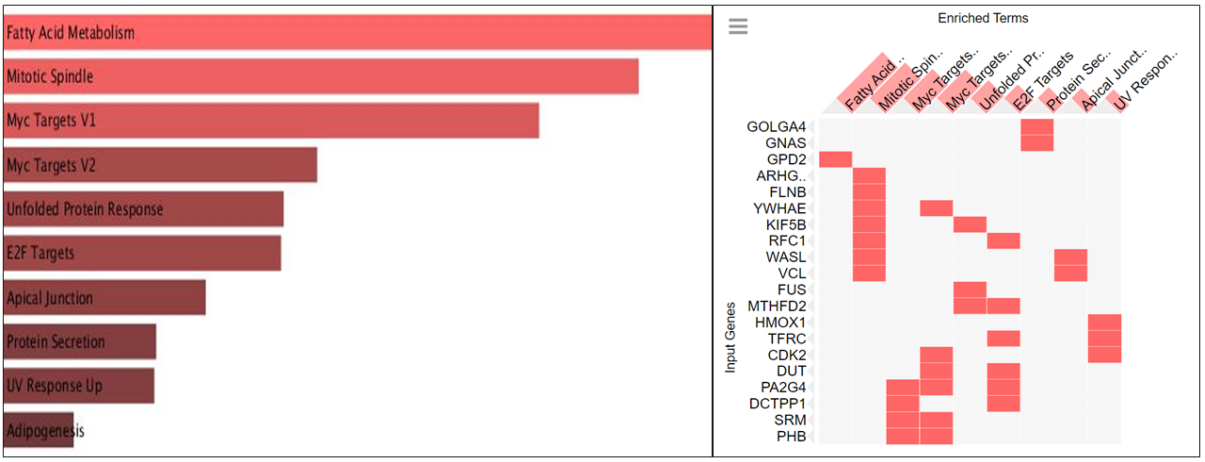
4.2. Enrichment Analysis of Final Gene Sets

Initially, an enrichment analysis was performed on all 363 genes amassed from the bootstrap iterations. The three principal enriched terms identified from the Reactome database (Figure 10) have all been correlated with breast cancer. RNA metabolism is essential in hormone signalling, and specific proteins and RNAs involved could serve as therapeutic targets [53]. Concerning the pathways involved in mRNA splicing, cancer cells have a considerably higher incidence of alternative mRNA splicing variants than non-cancer cells. Cancer cells are more sensitive to drugs targeting the splicing regulatory network. These findings suggest that modifying mRNA splicing has enormous potential as an approach for developing tailored therapeutics for breast cancer [54]. These pathways share the enrichment of the same genes, as observed in Figure 10(b). Considering the enriched pathways in the Hallmark dataset (Figure 11), there is a correlation between fatty acid metabolism and breast cancer. Breast cancer pathophysiology is tightly linked to fatty acid metabolism, and there is rising interest in using this metabolic route as a potential strategy for therapeutic techniques in breast cancer treatment [55]. Mitotic spindle microtubules are a significant target for breast cancer treatment [56]. There was no notable intersection between the gene sets of the three enriched pathways, as seen in Figure 11(b).



(a) (b)

Figure 10. Pathways enriched in Reactome database by analysing all 363 genes collected after bootstrap (a) Ranking of various pathways. (b) Grid analysis of genes associated with the pathways.

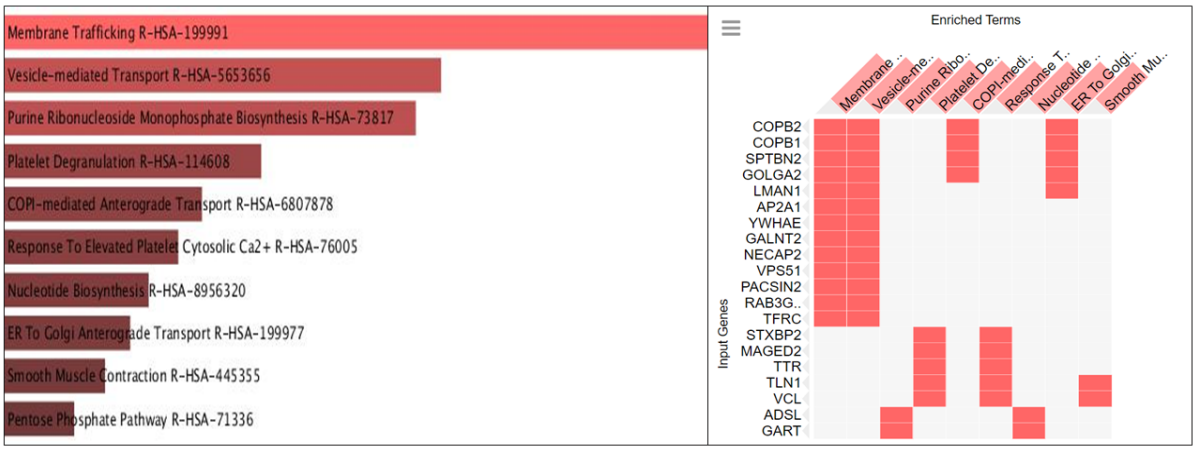


(a) (b)

Figure 11. Pathways enriched in Hallmark database by analysing all 363 genes collected after bootstrap (a) Ranking of various pathways. (b) Grid analysis of genes associated with the pathways.

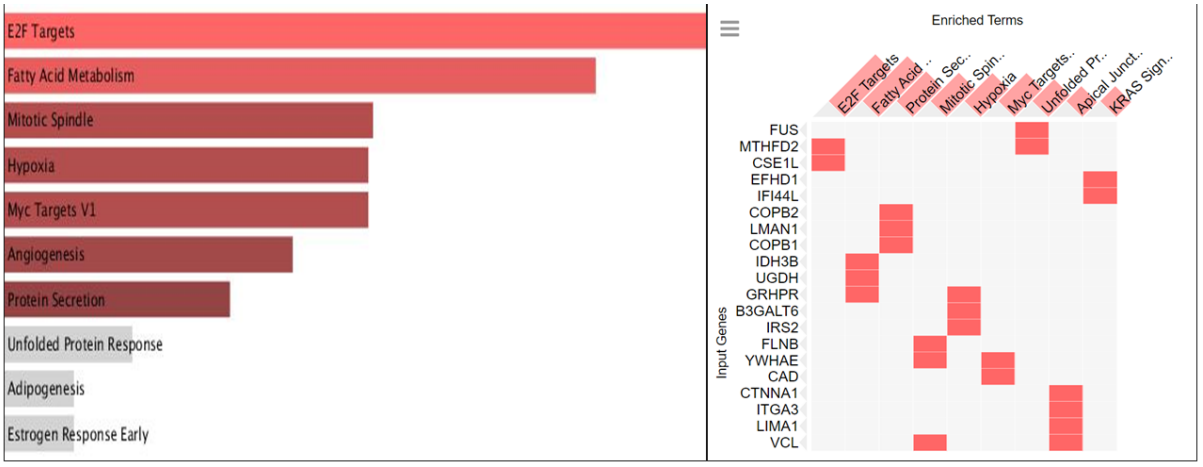
Following this, a filtration method was applied based on the frequency of gene recurrence over the iterations. A total of 168 genes were obtained after excluding those with a frequency below 100, refining the dataset. The subsequent pathway enrichment analysis was performed using the cohort consisting of 168 genes. Among the enriched pathways in the Reactome database (Figure 12), membrane trafficking pathways have been investigated as cancer therapeutic targets since they are crucial in various cellular processes. Protein kinase D (PKD) is involved in the secretory route at the trans-Golgi network, which is part of the membrane trafficking pathway, in breast cancer. The abnormal expression of PKD isoforms has been discovered mainly in breast cancer, indicating a possible therapeutic target [57]. Breast cancer development and metastasis have been linked to extracellular vesicle-mediated transport. Targeting proteins involved in vesicular transport may contribute to developing more effective therapeutic options for breast cancer therapy [58]. Membrane trafficking and vesicle-mediated transport have the same genes enriched. Platelets' potential in breast cancer targeted therapy has been highlighted by research: techniques targeting platelet receptors and using nano-platelets for drug delivery are

being developed. Furthermore, the influence of platelet degranulation on angiogenesis and cancer growth offers therapeutic promise in either suppressing cancer progression or facilitating targeted drug delivery [59,60]. The E2F pathway, essential for cell cycle progression, plays a vital role in the development of breast cancer, making it a prospective therapeutic target. Notably, in breast cancer targeted therapy, targeting the RB-E2F pathway, particularly by CDK inhibition (e.g., palbociclib), has emerged as a feasible method. Furthermore, specific E2F family members and pathway scores have been discovered as potential precision treatment targets and predictive biomarkers, which can help predict therapeutic response and patient prognosis. In addition, continuing research investigates genome-wide screens to improve the efficacy of combination medicines targeting E2F pathways in treating breast cancer [61,62]. Fatty acid metabolism, mitotic spindle and Myc targets v1 are still enriched pathways consistent with the observations from Figure 11.



(a) (b)

Figure 12. Pathways enriched in Reactome database by analysing genes with frequency > 100 (168 genes) (a) Ranking of various pathways. (b) Grid analysis of genes associated with the pathways.



(a) (b)

Figure 13. Pathways enriched in Hallmark database by analysing all genes with frequency > 100 (168 genes) (a) Ranking of various pathways. (b) Grid analysis of genes associated with the pathways.

By eliminating genes with a frequency greater than 150, we achieved an additional refinement of the gene set, which comprised a subset of 50 genes. Following this, an enrichment analysis was performed using the refined set of genes on the Reactome database. As shown in Figure 14, most of the enriched pathways are associated with breast cancer. IRS proteins, which have been linked to the development of breast cancer, regulate tumour cell survival and proliferation. We can observe in Figure 14(b) that the pathway is enriched due to IRS2. Although the connection between IRS activation and focused therapeutic techniques in breast cancer is not described in depth, it indicates a possible route for creating innovative treatment approaches [63,64]. An association has been established between the signalling mediated by SOS and the signalling mediated by IRS and breast cancer. It can be justified by Figure 14(b) as IRS2 is the gene behind the enrichment of the pathway. Regarding the observations obtained from the Hallmark database results, it is apparent that there is a significant enrichment in the Unfolded Protein Response (UPR) and Estrogen Response Late. Although both pathways have been linked to breast cancer, their relationships are distinct. Endocrine therapies that target oestrogen signalling are critical for treating ER+ breast cancer, which accounts for roughly 80% of all cases. [65,66,67]. The Unfolded Protein Response (UPR) pathway is a prospective target for breast cancer treatment since it is critical for controlling endoplasmic reticulum stress and protein homoeostasis. Its activation in tumour cells, which frequently have an increased protein burden, is linked to breast cancer growth, medication resistance, and poor survival outcomes. Targeting UPR and other protein quality control mechanisms is envisioned as a complete strategy for combating breast cancer, addressing issues such as recurrence, metastasis, and treatment resistance [68,69].

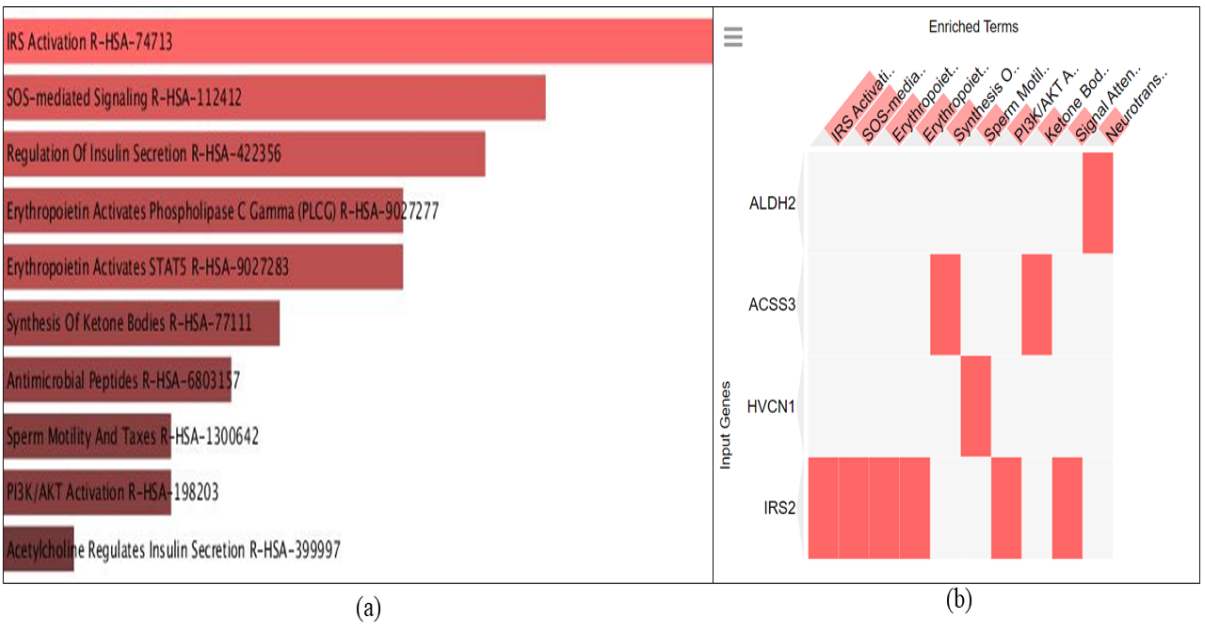
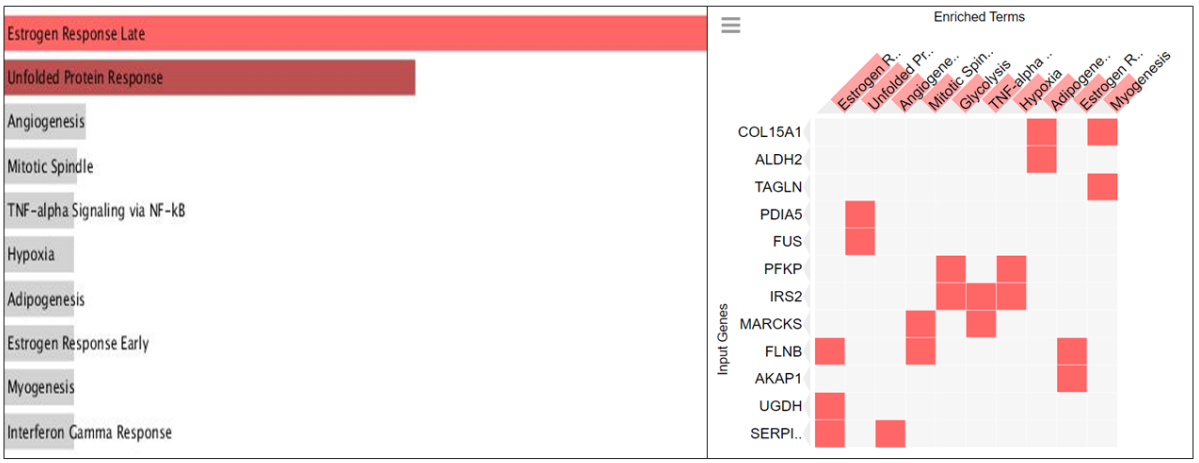
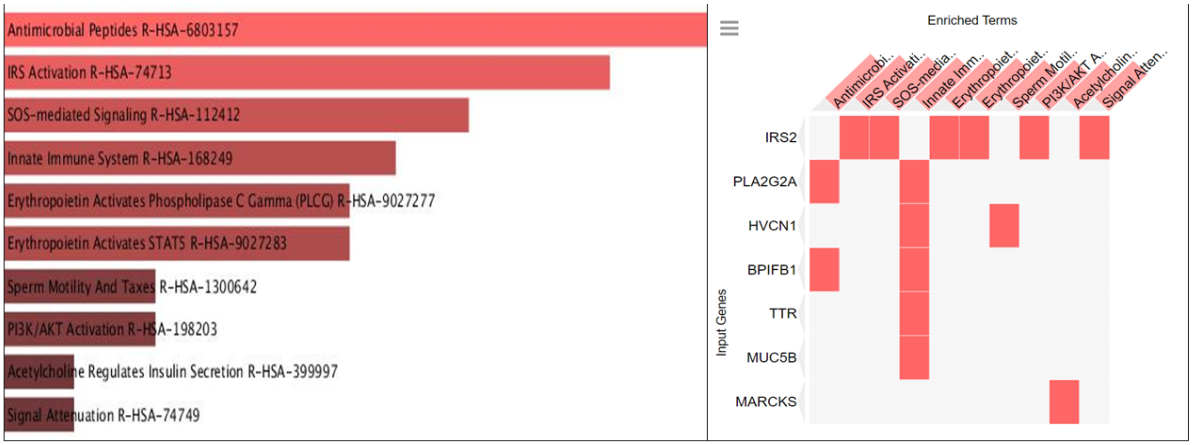


Figure 14. Pathways enriched in Reactome database by analysing all genes with frequency > 150 (50 genes) (a) Ranking of various pathways. (b) Grid analysis of genes associated with the pathways.



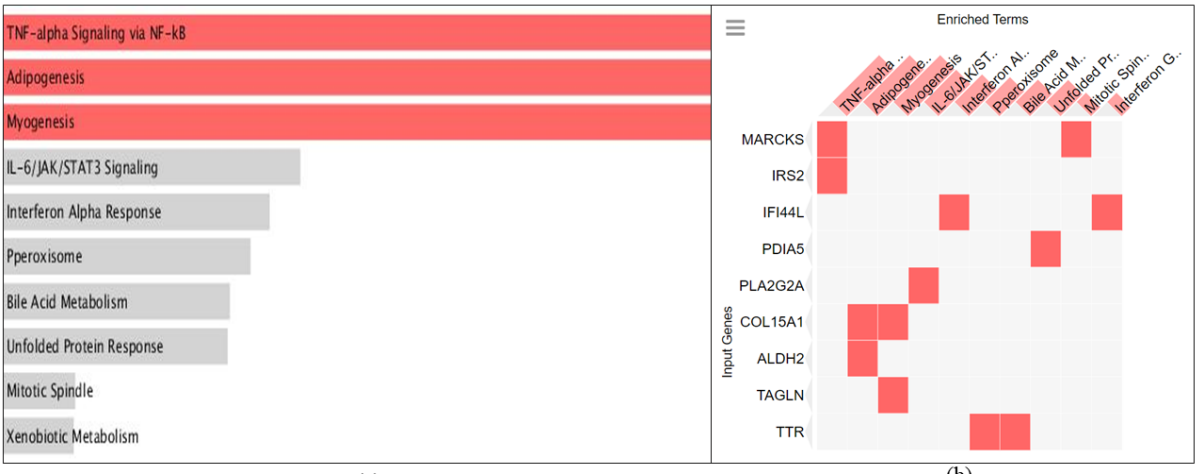
(a) (b)

Figure 15. Pathways enriched in Hallmark database by analysing all genes with frequency > 150 (50 genes) (a) Ranking of various pathways. (b) Grid analysis of genes associated with the pathways.



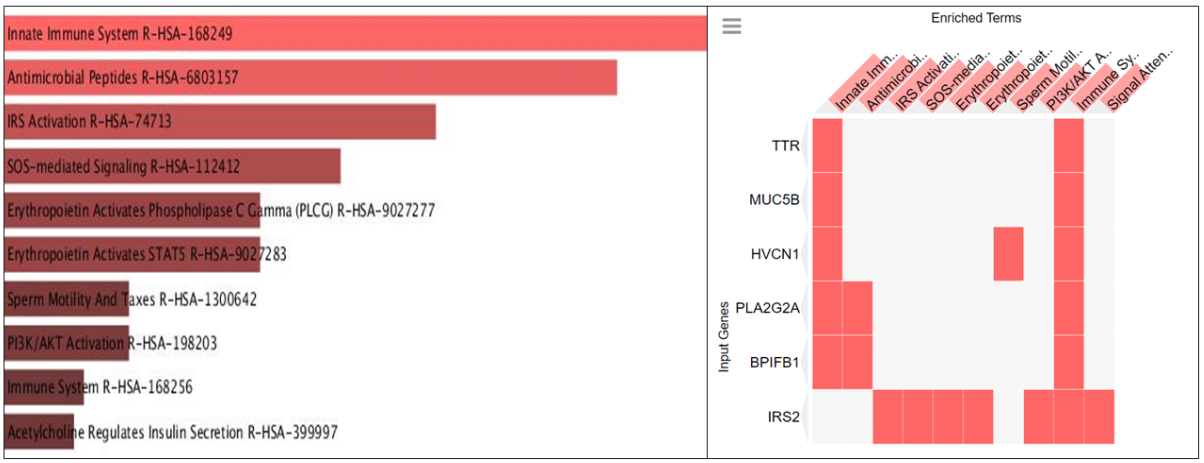
(a) (b)

Figure 16. Enriched pathways in the Reactome database by analysing the 25 most frequent genes (a) Ranking of various pathways. (b) Grid analysis of genes associated with the pathways.



(a) (b)

Figure 17. Enriched pathways in the Hallmark database by analysing the 25 most frequent genes (a) Ranking of various pathways. (b) Grid analysis of genes associated with the pathways.

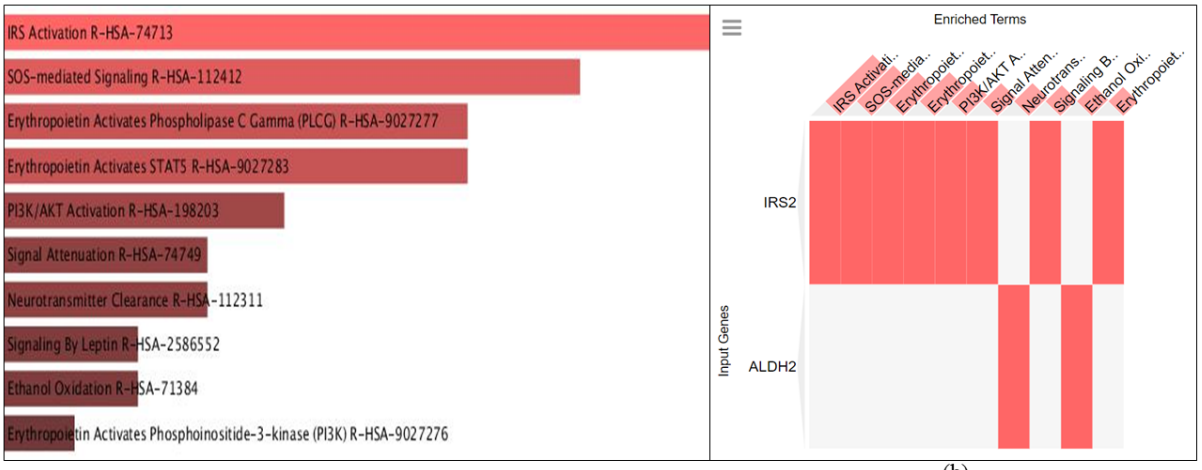


(a) (b)

Figure 18. Enriched pathways in the Reactome database by analysing the 20 most frequent genes (a) Ranking of various pathways. (b) Grid analysis of genes associated with the pathways.

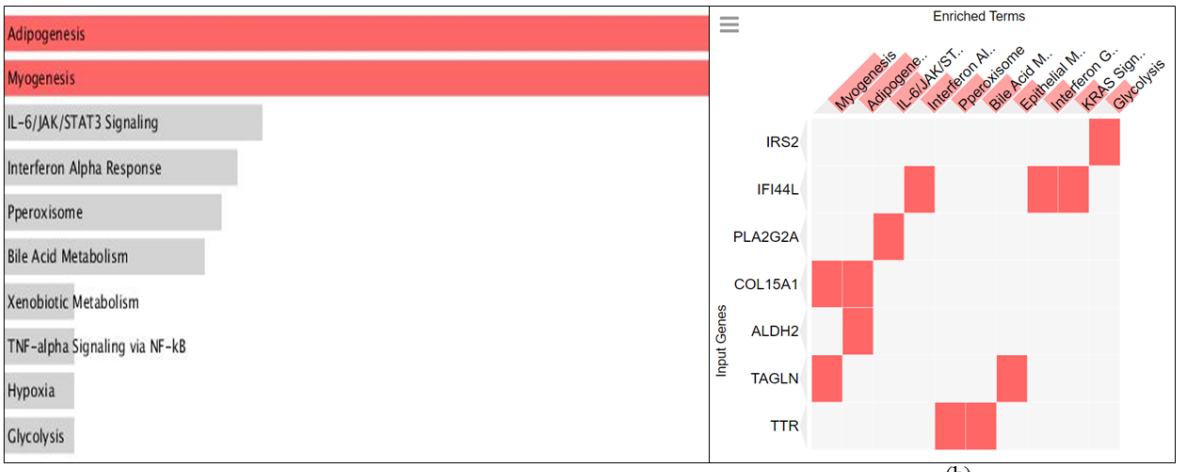
The gene set was then randomly reduced to subsets containing 25, 20, and 10 genes to assess the effect on the enriched pathways. After examining the 25 most frequently occurring genes in the Reactome dataset (Figure 16), it was observed that the Antimicrobial Peptides pathway displayed a greater degree of enrichment than prevalent pathways such as IRS Activation and SOS-mediated signalling. Antimicrobial peptides have garnered attention for their potential as targeted therapies for breast cancer [70,71]. Their ability to regulate inflammatory responses contributes to their recognition as a treatment option for breast cancer. The potential of utilising innate immune system targeting as a therapeutic frontier for breast cancer is emerging as a promising prospect [72,73]. A significant modification was identified in the enriched pathways within the Hallmark database (Figure 17). The TNF-alpha signalling is significant in breast cancer, but there has been no notable result for target therapy for the same [74]. Preliminary investigations on targeted therapy for the pathways involved in adipogenesis and myogenesis indicate promise, particularly when paired with traditional therapeutic methods, necessitating more study and clinical trials for validation but lacking conclusive evidence. The gene IRS2 is common in both databases' enriched pathways.

Removing five more genes based on frequency, it was observed that the rankings of enrichment in the Reactome pathways changed, as shown in Figure 18. The enrichment rankings for Hallmark pathways remained the same as in Figure 17. On examining the 15 most prevalent genes, it is observed that the IRS Activation pathway, SOS-mediated signalling pathway, and the Erythropoietin pathways emerged as the most enriched within the Reactome database (Figure 19). As illustrated in Figure 19, they were accompanied by the PI3K/AKT activation pathway. Because of its function in cancer cell proliferation and survival, the PI3K/AKT pathway is a prominent focus for targeted therapies in breast cancer carcinogenesis. Targeted therapies attempt to slow breast cancer growth by blocking this system's critical components. Targeting this route, particularly in aggressive subtypes like triple-negative breast cancer, has promise for improved therapy results [75,76,77]. IRS2 is the gene causing the enrichment of all the pathways. Within the Hallmark database, Adipogenesis and Myogenesis were the most enriched pathways (Figure 20).



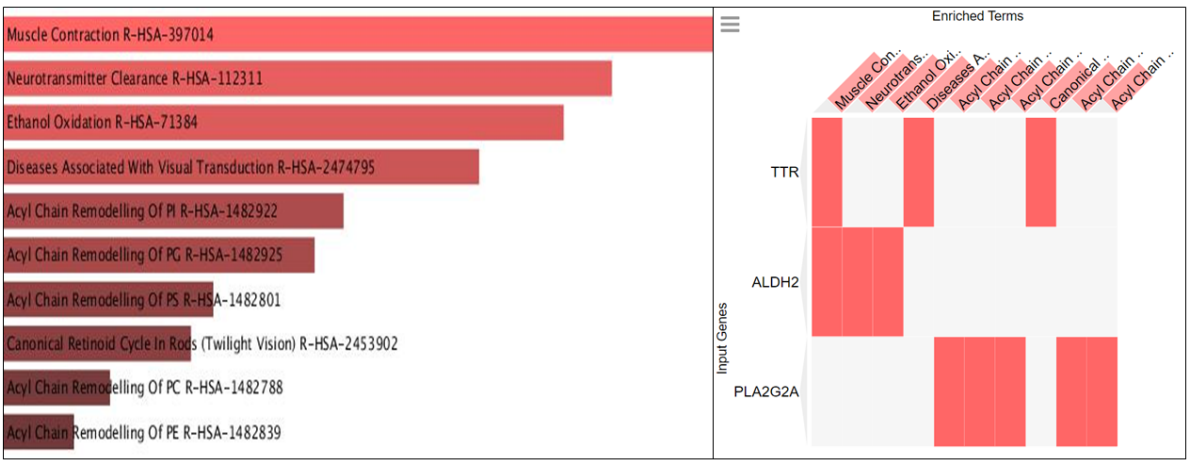
(a) (b)

Figure 19. Enriched pathways in the Reactome database by analysing the 15 most frequent genes (a) Ranking of various pathways. (b) Grid analysis of genes associated with the pathways.



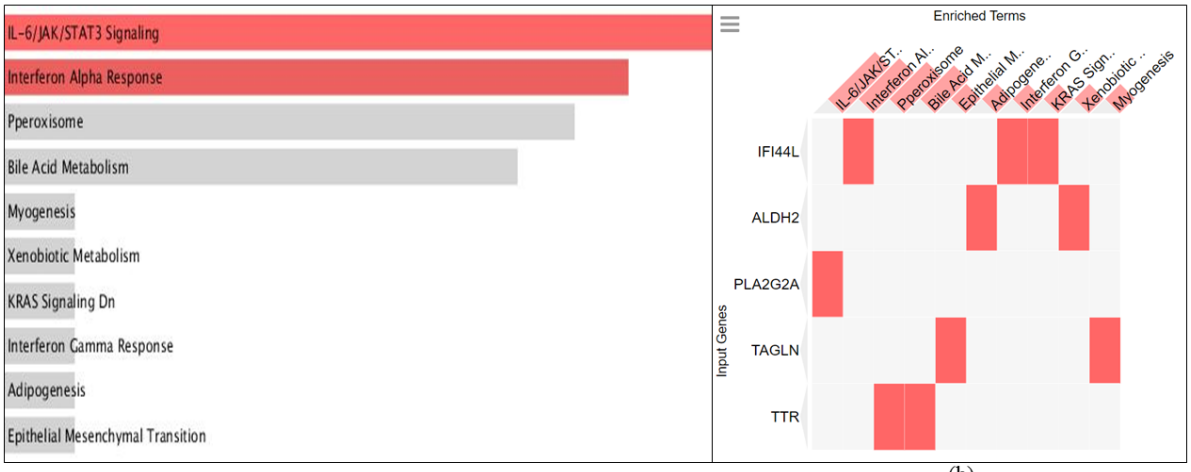
(a) (b)

Figure 20. Enriched pathways in the Hallmark database by analysing the 15 most frequent genes (a) Ranking of various pathways. (b) Grid analysis of genes associated with the pathways.



(a) (b)

Figure 21. Enriched pathways in the Reactome database by analysing the 10 most frequent genes (a) Ranking of various pathways. (b) Grid analysis of genes associated with the pathways.



(a) (b)

Figure 22. Enriched pathways in the Hallmark database by analysing the 10 most frequent genes (a) Ranking of various pathways. (b) Grid analysis of genes associated with the pathways.

Transitioning to the analysis of the top 10 genes, it is observed that the pathways of Muscle Contraction, Neurotransmitter Clearance, and Ethanol Oxidation are the most enriched in association with breast cancer for the Reactome database, as per Figure 21. No target therapy is aimed at the muscle contraction pathway in breast cancer. There have been documented correlations between breast cancer and exposure to ethanol and neurotransmitter clearance, but it is a potential area for further research [84,85,86]. The pathway enrichment observations for the Hallmark database are shown in Figure 22. The IL-6/JAK/STAT3 signalling system is critical in breast cancer development, particularly metastasis, making it an important therapeutic target. It has an important role in cancer cell proliferation, invasiveness, and apoptosis suppression, all contributing to metastasis. This mechanism acts independently of traditional ER-targeted medicines, necessitating alternate treatments. Several pathway components are being investigated as therapeutic targets, with developing monoclonal antibodies and medication combinations showing promise in preclinical and clinical studies to disrupt this system and improve breast cancer treatment results [78,79,80]. Interferon alpha (IFN) response and breast cancer have a noteworthy connection, highlighting the potential for targeted therapy. Depending on the transcribed interferon-stimulated genes, IFN signalling in breast cancer, particularly oestrogen receptor-positive and inflammatory subtypes, can contribute to anti-tumor responses and therapeutic resistance. Research on various drugs indicates that targeting the IFN pathway might be a potential option for targeted breast cancer therapy [81,82,83].

After analysing the grid clusters in each figure, it is observed that the genes PLA2G2A, BPIFB1 and IRS2 have been found in most of the pathways related to target therapy in breast cancer when treated with the AZD4547 FGFR inhibitor, as seen above.

5. LIMITATIONS AND FUTURE WORK

The study's restrictions are primarily defined by its theoretical and computational nature. The promising findings require further confirmation by empirical clinical studies to prove their dependability and practical application. The validity of the ML model is fundamentally related to the integrity of the training data; biases or mistakes in the data might alter predictions. Furthermore, the probable lack of comprehensive and

well-curated datasets may restrict the machine learning model's exhaustive validation. This shortcoming highlights an important concern: the model's limited ability to confirm its forecasting capabilities. The model's application may be limited to specific breast cancer subtypes. This specialisation raises concerns about the model's extrapolative capability and applicability to various breast cancer presentations. The expansion of the present dataset is an important step forward. The predicted accuracy and resilience of the model will be greatly improved by including a more extensive and heterogeneous data assortment.

The biomarkers identified in this work, while promising for FGFR-targeted treatment, are still in the early phases of development. Validation of developed biomarkers necessitates extensive in-vitro and in-vivo testing. These comprehensive investigations are critical in determining the effectiveness and safety of biomarkers in the context of FGFR-targeted treatment. Such empirical data will support the clinical feasibility of biomarkers, opening the way for personalised, effective, and safe treatment approaches for breast cancer patients globally.

6. CONCLUSION

This study conducted a detailed examination to provide important insights into the molecular basis of breast cancer. Numerous possible biomarkers and related pathways essential to breast cancer pathophysiology were found by methodically analysing a comprehensive pan-cancer dataset using a robust machine-learning framework. Significant genes associated with breast cancer growth and treatment response were identified using iterative feature selection strategies.

Enrichment analysis on gene sets using Enrichr utilising the Reactome and Hallmark pathway databases identified enriched pathways associated with breast cancer, such as Regulation of Expression of SLITs and ROBOs, Myc Targets V1, and Oestrogen Response. Notably, genes such as IRS2, PLA2G2A, and BPIFB1 were identified as involved in several pathways, emphasising their potential relevance in breast cancer pathophysiology. The downstream analysis gave a more nuanced knowledge of gene recurrence frequency throughout bootstrap iterations, assisting in the refinement of the gene set and revealing a variety of therapeutic targets.

This methodological approach, which combines bioinformatics and machine learning, was a big step towards a better understanding of the molecular complexity of breast cancer. The discovered genes and pathways add to our present molecular knowledge of breast cancer and enable future research into their potential as biomarkers or therapeutic targets.

The complete study calls for the ongoing development of data analysis approaches and stronger integrative frameworks to identify molecular complexities and therapeutic targets in breast cancer and beyond. This journey from data collecting to insightful findings is a big step forward in gaining a more nuanced understanding of breast cancer at the molecular level.

7. ACKNOWLEDGMENT

I wish to express my appreciation to my supervisor, Associate Professor Lan K. Nguyen, for his constant support, useful assistance, and insightful critiques throughout this project. His knowledge and guidance were invaluable in crafting my thesis and have greatly expanded my academic experience.

I am also grateful to my co-supervisor, Dr. Sungyoung Shin, whose constructive input, patience, and encouragement were invaluable in overcoming the project's hurdles. His broad expertise and unique ideas in bioinformatics

and machine learning have greatly contributed to the research.

I would also acknowledge the use of ChatGPT [87] to:

a. Validate the features provided by the machine learning framework in order to generate credible references supporting my work.
b. To improve the academic language and organisation of my own thesis work. This was further adjusted to fit my unique writing style.

Finally, I am grateful to The Faculty of Information Technology at Monash University for creating a welcoming and dynamic atmosphere for learning and enquiry, which greatly facilitated my research.

8. REFERENCES

[1] Perou, C. M., Sørlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., ... & Botstein, D. (2000). Molecular portraits of human breast tumours. *nature*, *406*(6797), 747-752.

[2] Ely, S., & Vioral, A. N. (2007). Breast Cancer Overview. *Plastic Surgical Nursing*, *27*(3), 128–133. https://doi.org/10.1097/01.psn.0000290281.48197.ae

[3] *Breast Cancer Biomarkers | ARUP Consult*. (n.d.). Arupconsult.com. https://arupconsult.com/content/breast-cancer

[4] Mohamed, A., Krajewski, K., Cakar, B., & Ma, C. X. (2013). Targeted Therapy for Breast Cancer. *The American Journal of Pathology*, *183*(4), 1096–1112. https://doi.org/10.1016/j.ajpath.2013.07.005

[5] Waks, A. G., & Winer, E. P. (2019). Breast Cancer Treatment. *JAMA*, *321*(3), 316. https://doi.org/10.1001/jama.2018.20751

[6] Higgins, M. J., & Baselga, J. (2011). Targeted therapies for breast cancer. *Journal of Clinical Investigation*, *121*(10), 3797–3803. https://doi.org/10.1172/jci57152

[7] Weigel, M. T., & Dowsett, M. (2010). Current and emerging biomarkers in breast cancer: prognosis and prediction. *Endocrine-Related Cancer*, *17*(4), R245–R262. https://doi.org/10.1677/erc-10-0136

[8] Shin, S.-Y., Centenera, M. M., Hodgson, J. T., Nguyen, E. V., Butler, L. M., Daly, R. J., & Nguyen, L. K. (2023). A Boolean-based machine learning framework identifies predictive biomarkers of HSP90-targeted therapy response in prostate cancer. *Frontiers in Molecular Biosciences*, *10*. https://doi.org/10.3389/fmolb.2023.1094321

[9] Adam, G., Rampášek, L., Safikhani, Z., Smirnov, P., Haibe-Kains, B., & Goldenberg, A. (2020). Machine learning approaches to drug response prediction: challenges and recent progress. *Npj Precision Oncology*, *4*(1). https://doi.org/10.1038/s41698-020-0122-1

[10] Nalejska, E., Mączyńska, E., & Lewandowska, M. A. (2014). Prognostic and Predictive Biomarkers: Tools in Personalized Oncology. *Molecular Diagnosis & Therapy*, *18*(3), 273–284. https://doi.org/10.1007/s40291-013-0077-9

[11] Cancer.net. (2019, June 11). *Breast Cancer - Diagnosis*. Cancer.net. https://www.cancer.net/cancer-types/breast-cancer/diagnosis

[12] Cleveland Clinic. (2022, January 21). *Breast Cancer: Causes, Stage, Diagnosis & Treatment*. Cleveland Clinic. https://my.clevelandclinic.org/health/diseases/3986-breast-cancer

[13] mayo clinic. (2019). *Breast cancer - Diagnosis and treatment - Mayo Clinic*. Mayoclinic.org. https://www.mayoclinic.org/diseases-condition

s/breast-cancer/diagnosis-treatment/drc-20352475

[14] *Tests to diagnose breast cancer | Cancer Research UK*. (2017). Cancerresearchuk.org. https://www.cancerresearchuk.org/about-cancer/breast-cancer/getting-diagnosed/tests-diagnose

[15] Diaz-Uriarte, R., Gómez de Lope, E., Giugno, R., Fröhlich, H., Nazarov, P. V., Nepomuceno-Chamorro, I. A., Rauschenberger, A., & Glaab, E. (2022). Ten quick tips for biomarker discovery and validation analyses using machine learning. *PLOS Computational Biology*, *18*(8), e1010357. https://doi.org/10.1371/journal.pcbi.1010357

[16] Patani, N., Martin, L.-A., & Dowsett, M. (2013). Biomarkers for the clinical management of breast cancer: International perspective. *International Journal of Cancer*, *133*(1), 1–13. https://doi.org/10.1002/ijc.27997

[17] Jhan, J.-R., & Andrechek, E. R. (2017). Triple-negative breast cancer and the potential for targeted therapy. *Pharmacogenomics*, *18*(17), 1595–1609. https://doi.org/10.2217/pgs-2017-0117

[18] Touat, M., Ileana, E., Postel-Vinay, S., André, F., & Soria, J.-C. (2015). Targeting FGFR Signaling in Cancer. *Clinical Cancer Research*, *21*(12), 2684–2694. https://doi.org/10.1158/1078-0432.CCR-14-2329

[19] Ye, T., Wei, X., Yin, T., Xia, Y., Li, D., Shao, B., Song, X., He, S., Luo, M., Gao, X., He, Z., Luo, C., Xiong, Y., Wang, N., Zeng, J., Zhao, L., Shen, G., Xie, Y., Yu, L., & Wei, Y. (2014). Inhibition of FGFR signaling by PD173074 improves antitumor immunity and impairs breast cancer metastasis. *Breast Cancer Research and Treatment*, *143*(3), 435–446. https://doi.org/10.1007/s10549-013-2829-y

[20] André, F., & Cortés, J. (2015). Rationale for targeting fibroblast growth factor receptor signaling in breast cancer. *Breast Cancer Research and Treatment*, *150*(1), 1–8. https://doi.org/10.1007/s10549-015-3301-y

[21] Mohamed, A., Krajewski, K., Cakar, B., & Ma, C. X. (2013). Targeted Therapy for Breast Cancer. *The American Journal of Pathology*, *183*(4), 1096–1112. https://doi.org/10.1016/j.ajpath.2013.07.005

[22] Jhan, J.-R., & Andrechek, E. R. (2017). Triple-negative breast cancer and the potential for targeted therapy. *Pharmacogenomics*, *18*(17), 1595–1609. https://doi.org/10.2217/pgs-2017-0117

[23] Dienstmann, R., Rodon, J., Prat, A., Perez-Garcia, J., Adamo, B., Felip, E., Cortes, J., Iafrate, A. J., Nuciforo, P., & Tabernero, J. (2014). Genomic aberrations in the FGFR pathway: opportunities for targeted therapies in solid tumors. *Annals of Oncology*, *25*(3), 552–563. https://doi.org/10.1093/annonc/mdt419

[24] Sobhani, N., Ianza, A., D'Angelo, A., Roviello, G., Giudici, F., Bortul, M., Zanconati, F., Bottin, C., & Generali, D. (2018). Current Status of Fibroblast Growth Factor Receptor-Targeted Therapies in Breast Cancer. *Cells*, *7*(7), 76. https://doi.org/10.3390/cells7070076

[25] Xie, Y., Su, N., Yang, J., Tan, Q., Huang, S., Jin, M., Ni, Z., Zhang, B., Zhang, D., Luo, F., Chen, H., Sun, X., Feng, J. Q., Qi, H., & Chen, L. (2020). FGF/FGFR signaling in health and disease. *Signal Transduction and Targeted Therapy*, *5*(1). https://doi.org/10.1038/s41392-020-00222-7

[26] Babina, I. S., & Turner, N. C. (2017). Advances and challenges in targeting FGFR signalling in cancer. *Nature Reviews Cancer*, *17*(5), 318–332. https://doi.org/10.1038/nrc.2017.8

[27] Chae, Y. K., Hong, F., Vaklavas, C., Cheng, H. H., Hammerman, P., Mitchell, E. P., Zwiebel, J. A., Ivy, S. P., Gray, R. J., Li, S., McShane, L. M., Rubinstein, L. V., Patton, D., Williams, P. M., Hamilton, S. R., Mansfield, A., Conley, B. A., Arteaga, C. L., Harris, L. N., & O'Dwyer, P. J. (2020). Phase II Study of AZD4547 in Patients With Tumors Harboring Aberrations in the FGFR Pathway: Results From the NCI-MATCH Trial (EAY131) Subprotocol W. *Journal of Clinical Oncology*, *38*(21), 2407–2417. https://doi.org/10.1200/jco.19.02630

[28] De Luca, A., Frezzetti, D., Gallo, M., & Normanno, N. (2017). FGFR-targeted therapeutics for the treatment of breast cancer. *Expert Opinion on Investigational Drugs*, *26*(3), 303–311. https://doi.org/10.1080/13543784.2017.1287173

[29] André, F., Bachelot, T., Campone, M., Dalenc, F., Perez-Garcia, J. M., Hurvitz, S. A., Turner, N., Rugo, H., Smith, J. W., Deudon, S., Shi, M., Zhang, Y., Kay, A., Porta, D. G., Yovine, A., & Baselga, J. (2013). Targeting FGFR with dovitinib (TKI258): preclinical and clinical data in breast cancer. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *19*(13), 3693–3702. https://doi.org/10.1158/1078-0432.CCR-13-0190

[30] *Gygi Lab @ HMS*. (n.d.). Gygi.hms.harvard.edu. Retrieved April 24, 2023, from https://gygi.hms.harvard.edu/publications/ccle.html

[31] *DepMap: The Cancer Dependency Map Project at Broad Institute*. (n.d.). Depmap.org. https://depmap.org/portal/

[32] *AZD4547 DepMap Compound Summary*. (n.d.). Depmap.org. Retrieved October 10, 2023, from https://depmap.org/portal/compound/AZD4547?tab=dose-curves

[33] Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G., Clark, N. R., & Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, *14*(1), 128. https://doi.org/10.1186/1471-2105-14-128

[34] Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., & Ma'ayan, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, *44*(W1), W90–W97. https://doi.org/10.1093/nar/gkw377

[35] Xie, Z., Bailey, A., Kuleshov, M. V., Clarke, D. J. B., Evangelista, J. E., Jenkins, S. L., Lachmann, A., Wojciechowicz, M. L., Kropiwnicki, E., Jagodnik, K. M., Jeon, M., & Ma'ayan, A. (2021). Gene Set Knowledge Discovery with Enrichr. *Current Protocols*, *1*(3). https://doi.org/10.1002/cpz1.90

[36] Elliot. (2021, March 3). *Another feature selection algorithm: MRMR*. Medium. https://elliot-weissberg.medium.com/another-feature-selection-algorithm-mrmr-3827b6b19e33

[37] smazzanti. (2023, October 3). *smazzanti/mrmr*. GitHub. https://github.com/smazzanti/mrmr

[38] Nusinow, D. P., & Gygi, S. P. (2020). A Guide to the Quantitative Proteomic Profiles of the Cancer Cell Line Encyclopedia. *BioRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2020.02.03.932384

[39] Waks, A. G., & Winer, E. P. (2019). Breast Cancer Treatment. *JAMA*, *321*(3), 316. https://doi.org/10.1001/jama.2018.20751

[40] Higgins, M. J., & Baselga, J. (2011). Targeted therapies for breast cancer. *Journal of Clinical Investigation*, *121*(10), 3797–3803. https://doi.org/10.1172/jci57152

[41] Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., & Mueller, A. (2015). Scikit-learn. *GetMobile: Mobile Computing and Communications*, *19*(1), 29–33. https://doi.org/10.1145/2786984.2786995

[42] *XGBoost Documentation — xgboost 1.6.1 documentation*. (n.d.). Xgboost.readthedocs.io. https://xgboost.readthedocs.io/en/stable/index.html

[43] Joshi, A., Van de Peer, Y., & Michoel, T. (2007). Analysis of a Gibbs sampler method for model-based clustering of gene expression data. *Bioinformatics*, *24*(2), 176–183. https://doi.org/10.1093/bioinformatics/btm562

[44] Sheng, Q., Gert Thijs, Moreau, Y., & Bart De Moor. (2006). Applications of Gibbs sampling in bioinformatics. *Optimization Methods & Software*.

[45] LeBlond, D. (2017). *Statistical Design and Analysis of Long-Term Stability Studies for Drug Products*. https://doi.org/10.1016/b978-0-12-802447-8.00022-4

[46] Gara, R. K., Kumari, S., Ganju, A., Yallapu, M. M., Jaggi, M., & Chauhan, S. C. (2015). Slit/Robo pathway: a promising therapeutic target for cancer. *Drug Discovery Today*, *20*(1), 156–164. https://doi.org/10.1016/j.drudis.2014.09.008

[47] Schulze, A., Oshi, M., Endo, I., & Takabe, K. (2020). MYC Targets Scores Are Associated with Cancer Aggressiveness and Poor Survival in ER-Positive Primary and Metastatic Breast Cancer. *International Journal of Molecular Sciences*, *21*(21), 8127. https://doi.org/10.3390/ijms21218127

[48] Wu, R., Sarkar, J., Yoshihisa Tokumaru, Yamato Takabe, Masanori Oshi, Asaoka, M., Yan, L., Ishikawa, T., & Kazuaki Takabe. (2022). Intratumoral lymphatic endothelial cell infiltration reflecting lymphangiogenesis is counterbalanced by immune responses and better cancer biology in the breast cancer tumor microenvironment. *PubMed*, *12*(2), 504–520.

[49] Tang, M., O'Grady, S., Crown, J., & Duffy, M. J. (2022). MYC as a therapeutic target for the treatment of triple-negative breast cancer: preclinical investigations with the novel MYC inhibitor, MYCi975. *Breast Cancer Research and Treatment*, *195*(2), 105–115. https://doi.org/10.1007/s10549-022-06673-6

[50] Wang, Y., Appiah-Kubi, K., Wu, M., Yao, X., Qian, H., Wu, Y., & Chen, Y. (2016). The platelet-derived growth factors (PDGFs) and their receptors (PDGFRs) are major players in oncogenesis, drug resistance, and attractive oncologic targets in cancer. *Growth Factors*, *34*(1-2), 64–71. https://doi.org/10.1080/08977194.2016.1180293

[51] Bonadonna, G., Valagussa, P., Tancini, G., & Di Fronzo, G. (1980). Estrogen-receptor status and response to chemotherapy in early and advanced breast cancer. *Cancer Chemotherapy and Pharmacology*, *4*(1), 37–41. https://doi.org/10.1007/BF00255456

[52] Ovaska, K., Matarese, F., Grote, K., Iryna Charapitsa, Cervera, A., Liu, C., Reid, G., Seifert, M., Stunnenberg, H. G., & Sampsa Hautaniemi. (2013). Integrative Analysis of Deep Sequencing Data Identifies Estrogen Receptor Early Response Genes and Links ATAD3B to Poor Survival in Breast Cancer. *PLOS Computational Biology*, *9*(6), e1003100–e1003100. https://doi.org/10.1371/journal.pcbi.1003100

[53] Wang, X., & Yang, D. (2021). The regulation of RNA metabolism in hormone

signaling and breast cancer. *Molecular and Cellular Endocrinology*, *529*, 111221. https://doi.org/10.1016/j.mce.2021.111221

[54] Fackenthal, J. D. (2023). Alternative mRNA Splicing and Promising Therapies in Cancer. *Biomolecules*, *13*(3), 561. https://doi.org/10.3390/biom13030561

[55] Munir, R., Lisec, J., Swinnen, J. V., & Zaidi, N. (2022). Too complex to fail? Targeting fatty acid metabolism for cancer therapy. *Progress in Lipid Research*, *85*, 101143. https://doi.org/10.1016/j.plipres.2021.101143

[56] Jordan, M. A., & Wilson, L. (2004). Microtubules as a target for anticancer drugs. *Nature Reviews Cancer*, *4*(4), 253–265. https://doi.org/10.1038/nrc1317

[57] Gutiérrez-Galindo, E., Zeynep Yılmaz, & Haußer, A. (2023). Membrane trafficking in breast cancer progression: protein kinase D comes into play. *Frontiers in Cell and Developmental Biology*, *11*. https://doi.org/10.3389/fcell.2023.1173387

[58] Mughees, M., Chugh, H., & Wajid, S. (2019). Vesicular trafficking–related proteins as the potential therapeutic target for breast cancer. *Protoplasma*, *257*(2), 345–352. https://doi.org/10.1007/s00709-019-01462-3

[59] Wang, L., Wang, X., Guo, E., Mao, X., & Miao, S. (2022). Emerging roles of platelets in cancer biology and their potential as therapeutic targets. *Frontiers in Oncology*, *12*. https://doi.org/10.3389/fonc.2022.939089

[60] Holmes, C. E., Levis, J. E., Schneider, D. J., Bambace, N. M., Sharma, D., Lal, I., Wood, M., & Muss, H. B. (2016). Platelet phenotype changes associated with breast cancer and its treatment. *Platelets*, *27*(7), 703–711. https://doi.org/10.3109/09537104.2016.1171302

[61] Oshi, M., Takahashi, H., Tokumaru, Y., Yan, L., Rashid, O. M., Nagahashi, M., Matsuyama, R., Endo, I., & Takabe, K. (2020). The E2F Pathway Score as a Predictive Biomarker of Response to Neoadjuvant Therapy in ER+/HER2− Breast Cancer. *Cells*, *9*(7), 1643. https://doi.org/10.3390/cells9071643

[62] Johnson, J., Thijssen, B., McDermott, U., Garnett, M., Wessels, L. F. A., & Bernards, R. (2016). Targeting the RB-E2F pathway in breast cancer. *Oncogene*, *35*(37), 4829–4835. https://doi.org/10.1038/onc.2016.32

[63] Ikink, G. J., Boer, M., Bakker, E. R. M., & Hilkens, J. (2016). IRS4 induces mammary tumorigenesis and confers resistance to HER2-targeted therapy through constitutive PI3K/AKT-pathway hyperactivation. *Nature Communications*, *7*(1). https://doi.org/10.1038/ncomms13567

[64] Gibson, S. L., Ma, Z., & Shaw, L. M. (2007). Divergent Roles for IRS-1 and IRS-2 in Breast Cancer Metastasis. *Cell Cycle*, *6*(6), 631–637. https://doi.org/10.4161/cc.6.6.3987

[65] Haines, Corinne N., Wardell, S. E., & McDonnell, Donald P. (2021). Current and emerging estrogen receptor-targeted therapies for the treatment of breast cancer. *Essays in Biochemistry*. https://doi.org/10.1042/ebc20200174

[66] Sannino, S., & Brodsky, J. L. (2017). Targeting protein quality control pathways in breast cancer. *BMC Biology*, *15*(1). https://doi.org/10.1186/s12915-017-0449-4

[67] Xu, D., Liu, Z., Liang, M.-X., Fei, Y.-J., Zhang, W., Wu, Y., & Tang, J.-H. (2022). Endoplasmic reticulum stress targeted therapy for breast cancer. *Cell Communication and Signaling*, *20*(1). https://doi.org/10.1186/s12964-022-00964-7

[68] Masanori Oshi, Arya Mariam Roy, Gandhi, S., Yoshihisa Tokumaru, Yan, L., Yamada, A., Endo, I., & Kazuaki Takabe. (2022). *The clinical relevance of unfolded protein response signaling in breast cancer*. https://doi.org/10.21203/rs.3.rs-1480002/v1

[69] Patra, A., Adhikary, A., & Ghosh, S. (2022). The unfolded protein response (UPR) pathway: the unsung hero in breast cancer management. *Apoptosis*. https://doi.org/10.1007/s10495-022-01803-3

[70] Freitas, E. D., Bataglioli, R. A., Oshodi, J., & Beppu, M. M. (2022). Antimicrobial peptides and their potential application in antiviral coating agents. *Colloids and Surfaces B: Biointerfaces*, *217*, 112693. https://doi.org/10.1016/j.colsurfb.2022.112693

[71] Sultana, A., Luo, H., & Ramakrishna, S. (2021). Antimicrobial Peptides and Their Applications in Biomedical Sector. *Antibiotics*, *10*(9), 1094. https://doi.org/10.3390/antibiotics10091094

[72] Ye, Y., Xu, C., Chen, F., Liu, Q., & Cheng, N. (2021). Targeting Innate Immunity in Breast Cancer Therapy: A Narrative Review. *Frontiers in Immunology*, *12*. https://doi.org/10.3389/fimmu.2021.771201

[73] Shihab, I., Khalil, B. A., Elemam, N. M., Hachim, I. Y., Hachim, M. Y., Hamoudi, R. A., & Maghazachi, A. A. (2020). Understanding the Role of Innate Immune Cells and Identifying Genes in Breast Cancer Microenvironment. *Cancers*, *12*(8), 2226. https://doi.org/10.3390/cancers12082226

[74] Mercogliano, M. F., Bruni, S., Elizalde, P. V., & Schillaci, R. (2020). Tumor Necrosis Factor α Blockade: An Opportunity to Tackle Breast Cancer. *Frontiers in Oncology*, *10*. https://doi.org/10.3389/fonc.2020.00584

[75] Cidado, J., & Park, B. H. (2012). Targeting the PI3K/Akt/mTOR Pathway for Breast Cancer Therapy. *Journal of Mammary Gland Biology and Neoplasia*, *17*(3-4), 205–216. https://doi.org/10.1007/s10911-012-9264-2

[76] Zhu, K., Wu, Y., He, P., Fan, Y., Zhong, X., Zheng, H., & Luo, T. (2022). PI3K/AKT/mTOR-Targeted Therapy for Breast Cancer. *Cells*, *11*(16), 2508. https://doi.org/10.3390/cells11162508

[77] Yuan, Y., Long, H., Zhou, Z., Fu, Y., & Jiang, B. (2023). PI3K–AKT-Targeting Breast Cancer Treatments: Natural Products and Synthetic Compounds. *Biomolecules*, *13*(1), 93. https://doi.org/10.3390/biom13010093

[78] Chen, J., Wei, Y., Yang, W., Huang, Q., Chen, Y., Zeng, K., & Chen, J. (2022). IL-6: The Link Between Inflammation, Immunity and Breast Cancer. *Frontiers in Oncology*, *12*, 903800. https://doi.org/10.3389/fonc.2022.903800

[79] Manore, S. G., Doheny, D. L., Wong, G. L., & Lo, H.-W. (2022). IL-6/JAK/STAT3 Signaling in Breast Cancer Metastasis: Biology and Treatment. *Frontiers in Oncology*, *12*. https://doi.org/10.3389/fonc.2022.866014

[80] Manore, S. G., Doheny, D. L., Wong, G. L., & Lo, H.-W. (2022). IL-6/JAK/STAT3 Signaling in Breast Cancer Metastasis: Biology and Treatment. *Frontiers in Oncology*, *12*. https://doi.org/10.3389/fonc.2022.866014

[81] Zhao, N., Kabotyanski, E. B., Saltzman, A. B., Malovannaya, A., Yuan, X., Reineke, L. C., Lieu, N., Gao, Y., Pedroza, D. A., Sebastián Calderón, Smith, A. J., Hamor, C., Safari, K., Savage, S. R., Zhang, B., Zhou, J., Solis, L. M., Hilsenbeck, S. G., Fan, C., & Perou, C. M. (2023). Targeting EIF4A triggers an interferon response to synergize with chemotherapy and suppress triple-negative breast cancer. *BioRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2023.09.28.559973

[82] Provance, O. K., & Lewis-Wambi, J. (2019). Deciphering the role of interferon alpha signaling and microenvironment crosstalk in inflammatory breast cancer. *Breast Cancer Research*, *21*(1). https://doi.org/10.1186/s13058-019-1140-1

[83] Provance, O. K., & Lewis-Wambi, J. (2019). Deciphering the role of interferon alpha signaling and microenvironment crosstalk in inflammatory breast cancer. *Breast Cancer Research*, *21*(1). https://doi.org/10.1186/s13058-019-1140-1

[84] Corsello, S. M., Nagari, R. T., Spangler, R. D., Rossen, J., Kocak, M., Bryan, J. G., Humeidi, R., Peck, D., Wu, X., Tang, A. A., Wang, V. M., Bender, S. A., Lemire, E., Narayan, R., Montgomery, P., Ben-David, U., Garvie, C. W., Chen, Y., Rees, M. G., & Lyons, N. J. (2020). Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nature Cancer*. https://doi.org/10.1038/s43018-019-0018-6

[85] Faizi, N., & Alvi, Y. (2023). Correlation. *Elsevier EBooks*, 109–126. https://doi.org/10.1016/b978-0-443-18550-2.00002-5

[86] *5.6 - The General Linear F-Test | STAT 462*. (n.d.). Online.stat.psu.edu. Retrieved October 17, 2023, from https://online.stat.psu.edu/stat462/node/135/

[87] OpenAI. (2023). *ChatGPT* (Mar 14 version) [Large language model]. https://chat.openai.com/chat

PART III: APPENDIX

1. SOFTWARE AND ALGORITHMS

Python	https://python.org/	v 3.11.4
Sci-kit-learn	https://scikit-learn.org/	v 1.3.1
Seaborn	https://seaborn.pydata.org/	v 0.12.2
XGBoost Regressor	https://xgboost.readthedocs.io/en/stable/index.html#	v 1.7.6
mRMR	https://github.com/smazzanti/mrmr	v 0.2.8
Scipy	https://scipy.org/	v 1.10.1
Pandas	https://pandas.pydata.org/	v 1.5.3
Numpy	https://numpy.org/	v 1.24.3
Matplotlib	https://matplotlib.org/	v 3.7.1

2. SOURCE CODE FOR MACHINE LEARNING PIPELINE

```
#  Importing all the necessary libraries for EDA
import pandas as pd
import numpy as np
import re
import seaborn as sns
import matplotlib.pyplot as plt
import random
import os
%matplotlib inline


# For mrmr algorithm
```

```
!pip install xgboost
!pip install mrmr_selection
from mrmr import mrmr_regression

df = pd.read_csv('./CancerCell2022_AZD4547_PRISM.csv')

df_num = df.rename(columns = {"Row" : "Cell line"})
df_num.head()

df_num.info()

df_num.shape

df_num.describe()

# For categorical variables
df_num.describe(include=['O'])

# Checking for columns with NaN values (Missing Values)
[features for features in df_num.columns if df_num[features].isna().sum()>0]

# Checking the number of missing values in AUC
df_num['AUC'].isna().sum()

# Deleting rows with NaN
df_num = df_num.dropna()
```

```
df_num.isna().sum()


# Checking for duplicate rows in the dataset
df_num.duplicated().any()


df_num.info()


df_num.drop(["Cell line"], axis = 1, inplace = True)


df_num.head()




from sklearn.utils import resample


def bootstrap_data(dataset):
        print("Bootstrapping dataset")
        return resample(dataset, replace = True, n_samples = len(dataset))


from sklearn.model_selection import KFold


from sklearn.preprocessing import StandardScaler


# Defining a function to normalize the dataset


def norm(data):
        """
```

```
        The function takes a pandas dataset as input and returns a normalized pandas dataset (using Standard Scaler) as output
        """
        print(">Normalizing dataset")
        # Initializing Standard Scaler for normalizing dataset
        scaler = StandardScaler()
        normalized_data = scaler.fit_transform(data)

    return pd.DataFrame(normalized_data, columns = data.columns)

from scipy.stats import pearsonr

def corr_pearson(X, y):
        """
        This function takes the dataframe and the target variable as the input arguments and forms a dictionary with the features as keys
        and the corresponding correlation and p values as values. It then converts the dictionary into a pandas dataframe and returns only
    those features with p_value < 0.05 as function output.
        """
        print(">Feature selection using Pearson correlation")

    corr_n_p = {}

        for column in X:
        # pearsonr returns a tuple (Pearson's correlation coefficient, 2-tailed p-value)
        corr, p_value = pearsonr(X[column], y)
        corr_n_p[column] = (corr, p_value)
```

```
        correlation_df = pd.DataFrame.from_dict(corr_n_p, orient='index',
columns=['correlation', 'p_value'])


    # Reset the index to move the current index (genes) as a proper column

        correlation_df.reset_index(inplace=True)

        # Rename the column to 'genes'

        correlation_df.rename(columns={'index': 'genes'}, inplace=True)


    # Plotting the graph of correlation vs p-value

        plt.figure(figsize=(4, 3))

        sns.scatterplot(data=correlation_df, x='correlation', y='p_value', alpha=0.6,
edgecolor=None)

        plt.axhline(y=0.05, color='r', linestyle='--')

        plt.title('Correlation vs P-Value')

        plt.xlabel('Correlation')

        plt.ylabel('P-Value')

        plt.show()


    # Filtering and sorting the significant features

        significant_df = correlation_df[correlation_df.p_value <
0.05].sort_values(by='p_value', ascending=True)


    significant_df.reset_index(inplace=True)

        significant_df.drop(columns="index", inplace=True)


    # Filtering out the significant features from the actual dataset

        features = list(significant_df.genes)


    return features
```

```
# Defining a function to implement mRMR feature selection
def mrmr_feature_selection(X,y,k):
        # the different column types in the dataset
        print(">Implementing MRMR feature selection")
        return mrmr_regression(X =X, y= y, K = k)

from sklearn.feature_selection import f_regression, SelectKBest

# Defining a function to implement F-regression
def f_reg_feature_selection(X, y, k):
        print(">Implementing F-regression for feature selection")
        # Using SelectKBest to select the best features
        selector = SelectKBest(f_regression, k = k)
        new = selector.fit_transform(X,y)

   f_selected = X.columns[selector.get_support()].to_list()

   return f_selected

from sklearn.linear_model import LassoCV
from sklearn.linear_model import Lasso

def lasso_cv(X,y):
        print(">Implementing Lasso regularization for feature selection")

   # Initializing an array of different values for alpha
        alphas = np.logspace(-4,4,50)
```

```
    # Ensure X and y have matching indices
        X = X.reset_index(drop=True)
        y = y.reset_index(drop=True)

    # Use LassoCV to find the best alpha using CV
        lassocv = LassoCV(alphas = alphas, cv = 5)
        lassocv.fit(X,y)

    best_alpha = lassocv.alpha_
        print(f" best alpha value: {best_alpha}")

    # Using the best alpha to implemene Lasso Regularization
        lasso = Lasso(alpha=best_alpha)

        # Fit the Lasso model
        lasso.fit(X, y)

        # Identify features with non-zero coefficients
        lasso_features = np.where(lasso.coef_ != 0)[0]

    return lasso_features

from sklearn.model_selection import cross_val_score, RandomizedSearchCV
from sklearn.svm import SVR
import xgboost as xgb

# XGBoost
def xgb_regressor(X,y):
```

```
xgb_results = {}

        print('Fitting XGBoost Model')

        # Create a XGBoost Regressor object

        xgbr = xgb.XGBRegressor(n_estimators=100, learning_rate=0.1, random_state=0,
n_jobs=-1)


    # Param grid for XGBoost

        param_grid_xgb = {

        'subsample': [0.6, 0.7, 0.8, 0.9],

        'reg_lambda': [0.2, 0.5, 0.8, 1, 1.2],

        'reg_alpha': [0, 0.2, 0.5, 0.8, 1],

        'n_estimators': [100, 150, 200, 250, 300],

        'min_child_weight': [1, 2, 3, 4],

        'max_depth': [5, 6, 7, 8, 9],

        'learning_rate': [0.01, 0.05, 0.1, 0.15, 0.2],

        'gamma': [0, 0.1, 0.2, 0.3, 0.4],

        'colsample_bytree': [0.6, 0.7, 0.8, 0.9]}



        model = xgb.XGBRegressor()


        kf = KFold(n_splits=6, shuffle=True, random_state=42)


        grid_search = RandomizedSearchCV(model, param_grid_xgb, cv=kf, verbose=1)


    grid_search.fit(X,y)


    print(f"XGBoost Best Parameters: {grid_search.best_params_}")
```

```
        print(f"XGBoost Best Score: {grid_search.best_score_}")

    xgb_results['best_params'] = grid_search.best_params_
        xgb_results['best_score'] = grid_search.best_score_

    return xgb_results

# SVM

def svmachine(X,y):

    svm_results = {}

    print('Fitting Support Vector Machine Model')

    # Param grid for SVM
        param_grid_svm = {
        'shrinking': [True, False],
    'kernel': ['rbf', 'poly', 'sigmoid', 'linear'],
        'gamma': [0.001, 0.005, 0.01, 0.05, 0.1, 0.5],
        'epsilon': [0.0001, 0.001, 0.005, 0.01, 0.05, 0.1],
        'degree': [2, 3, 4, 5, 6],
        'coef0': [-1, -0.5, 0, 0.5, 1, 1.5],
        'C': [0.1, 0.5, 1, 5, 10, 25, 50, 100]}

    model = SVR()

        kf = KFold(n_splits=6, shuffle=True, random_state=42)
```

```
        grid_search = RandomizedSearchCV(model, param_grid_svm, cv=kf, verbose=1, n_jobs=-1)

    grid_search.fit(X, y)

    print(f"SVM Best Parameters: {grid_search.best_params_}")
        print(f"SVM Best Score: {grid_search.best_score_}")

    svm_results['best_params'] = grid_search.best_params_
        svm_results['best_score'] = grid_search.best_score_

    return svm_results

# MLP (Neural Network)

from sklearn.neural_network import MLPRegressor

def mlp_regressor(X,y):

    mlp_results = {}

    # Param Grid for MLP
        param_grid_mlp = {
        'hidden_layer_sizes': [(50, 50), (100, 100), (50, 50, 50), (150, 150), (100, 100, 100), (50, 50, 50, 50)],
        'activation': ['logistic', 'tanh', 'relu', 'identity'],
        'solver': ['sgd', 'adam', 'lbfgs'],
        'alpha': [0.0001, 0.0005, 0.001, 0.005, 0.01, 0.1],
```

```
        'learning_rate': ['constant', 'adaptive', 'invscaling'],
        'learning_rate_init': [0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05],
        'max_iter': [50, 100, 200, 300, 400, 500],
        'momentum': [0.1, 0.5, 0.7, 0.9, 0.95],
        'beta_1': [0.7, 0.8, 0.9, 0.95],
        'beta_2': [0.99, 0.995, 0.999]}


    print('Fitting Multi-Layer Perceptron Model (Neural Network)')

    kf = KFold(n_splits=6, shuffle=True, random_state=42)

    # Initialize MLPRegressor and GridSearchCV
        mlp = MLPRegressor(max_iter=1000) # You might want to increase max_iter if the model doesn't converge
        grid_search = RandomizedSearchCV(mlp, param_grid_mlp, n_jobs=-1, cv=kf, verbose=1)

        # Fit the model
        grid_search.fit(X, y)

    print(f"MLP Best Parameters: {grid_search.best_params_}")
        print(f"MLP Best Score: {grid_search.best_score_}")

    mlp_results['best_params'] = grid_search.best_params_
        mlp_results['best_score'] = grid_search.best_score_

    return mlp_results
```

```
# Defining a function to fit models
def model_fitting(X, y):
        print(">Fitting models")

    results = {}
        scores = []
        model_names = []

    # Cross validation object
        kf = KFold(n_splits=6, shuffle=True, random_state=42)
        print('-- Fitting models on default parameters...')
        models = {
        'SVM': SVR(),
        'MLP': MLPRegressor(),
        'XGBoost': xgb.XGBRegressor()
        }

    for name, model in models.items():

       cv_score_list = cross_val_score(model, X, y, cv = kf)
        results[f'{name}'] = (np.mean(cv_score_list))

    # Selecting the models with the best scores on default values
        sorted_results = sorted(results.items(), key=lambda x: x[1], reverse=True)
        top_2_keys = [sorted_results[i][0] for i in range(2)]

    tuning_models = {}
```

```
print(f'--Top 2 models on default parameters:{top_2_keys}')
        print('-- Hyperparameter tuning the models')
        for mod in  top_2_keys:
        if mod == 'SVM':
       svm_results = svmachine(X, y)
        tuning_models['SVM'] = svm_results
        elif mod == 'MLP':
        mlp_results = mlp_regressor(X, y)
        tuning_models['MLP'] = mlp_results
        elif mod == 'XGBoost':
        xgb_results = xgb_regressor(X, y)
        tuning_models['XGBoost'] = xgb_results


tuning_results = {}
        scores = []
        model_names = []


for name, model in tuning_models.items():


    tuning_results[f'{name}_score'] = (model['best_score'], model['best_params'])
        # for plotting
        scores.append(model['best_score'])
        model_names.append(name)


# Plotting the scores
        plt.figure(figsize=(3, 2))
        plt.barh(model_names, scores, color=['blue', 'red', 'green', 'yellow'])
        plt.xlabel('Best Score')
```

```
        plt.title('Model Performance')
        plt.gca().invert_yaxis()  # To display the model with the lowest MSE at the top
        plt.show()

    print("....Model Fitting Done")
        top2 = sorted(tuning_results.items(), key=lambda x: x[1][0], reverse=True)[:2]
        print(top2)
        return top2

from sklearn.feature_selection import SequentialFeatureSelector
from sklearn.model_selection import cross_val_score

def forward_backward_selection(X, y, top2_models):
        """
        This function performs forward feature selection and backward feature elimination on the dataset
        for the top 2 models and returns the selected features for each model and method.
        """
        kf = KFold(n_splits=6, shuffle=True, random_state=42)

    # Initializing dictionary to store results
        final_features = {}
        results = {}
        forward_counts = []
        backward_counts = []
        model_names = []

    # Extracting the best estimator from the models' results
```

```
        best_model = top2_models[0][0].split('_')[0]
        print(best_model)

    for model_name, (score, params) in top2_models[:1]:
        if 'MLP' in model_name.upper():
        model = MLPRegressor(**params)

      elif 'SVM' in model_name.upper():
        model = SVR(**params)

        elif 'XGBOOST' in model_name.upper():
        model = xgb.XGBRegressor(**params)

      else:
        raise ValueError(f'Unknown Model name: {model_name}')

        # Forward Feature Selection
        print(f">Running Forward Feature Selection for {model_name}")

        # Forward Selection
        sfs_forward = SequentialFeatureSelector(model, direction='forward', cv=kf,
n_features_to_select=20, n_jobs=-1)
        sfs_forward.fit(X, y)
        forward_features = X.columns[sfs_forward.get_support()]
        print(f'Forward features:{forward_features}')

      print(f">Running Backward Feature Selection for {model_name}")
        # Backward Selection
```

```
        sfs_backward = SequentialFeatureSelector(model, direction='backward', cv=kf,
n_features_to_select=20, n_jobs=-1)

         sfs_backward.fit(X, y)

        backward_features = X.columns[sfs_backward.get_support()]

        print(f'Backward features:{backward_features}')



        # Store results

        results[model_name.split('_')[0] + "_forward"] = forward_features,

        results[model_name.split('_')[0] + "_backward"] = backward_features


    print(results)

        return results


def pipeline(dataset, n_bootstraps):

        kf = KFold(n_splits=6, shuffle=True, random_state=42)


    # DataFrame to store the results for each bootstrap iteration

        results_df = pd.DataFrame(columns=['Bootstrap Iteration Count', 'Best Model',

                            'Best Model Forward Features', 'Best Model Backward Features','Best
Model Score'])


    # Iterator for bootstraps

        count = 1


    # Create the csv file with headers if it doesn't exist

        csv_file = 'bootstrap_final.csv'


    if not os.path.exists(csv_file):
```

```
            results_df.to_csv(csv_file, index = False)


    for _ in range(n_bootstraps):
            print("Bootstrapped sample {} of {}".format(count, n_bootstraps))


        # Bootstrapping
            bootstrapped_data = bootstrap_data(dataset)


        # Train and test
            X = bootstrapped_data.drop(columns=['AUC'])
            y = bootstrapped_data['AUC']
            X = X.reset_index(drop=True)
            y = y.reset_index(drop=True)
            # Normalizing
            X_normalized = norm(X)


        # Feature Selection
            # Pearson
            pearson_features = corr_pearson(X_normalized, y)


        if len(pearson_features)>500:
            # MRMR
            mrmr_features = mrmr_feature_selection(X_normalized[pearson_features], y, 500)
            # F-regression
            f_features = f_reg_feature_selection(X_normalized[pearson_features], y, 500)
                # Common between mrmr and f-regression
            common_features = list(set(mrmr_features) & set(f_features))
            print("   Number of common features between mrmr and
f-regression:{}".format(len(common_features)))
```

```
        else:
            common_features = pearson_features

        common_features_df = pd.DataFrame(common_features)
            # Write the DataFrame to a CSV file
            common_features_df.to_csv('./filter_fs_df.csv', index=False)
            # Lasso
              embedded_features_indices = lasso_cv(X_normalized[common_features], y)
            embedded_features = X_normalized.columns[embedded_features_indices].to_list()

        embedded_df = pd.DataFrame(embedded_features)
            # Write the DataFrame to a CSV file
            embedded_df.to_csv('./embedded_df.csv', index=False)

        print("       Number of common features left after
Lasso:{}".format(len(embedded_features)))

        # Model Selection
            top2_models = model_fitting(X_normalized[embedded_features], y)



        # Extracting model names
            best_model_name = top2_models[0][0].split('_')[0]  # Extracting the model name
from the result tuple
            best_model_r2 = top2_models[0][1][0]
       print(f"       The best model is {best_model_name} with R2 score of {best_model_r2}")

        if len(embedded_features)>20: # We are trying to filter out the best 20 features
```

```
            # FFS, BFE
            final_features = forward_backward_selection(X_normalized[embedded_features], y, top2_models)

            # Extracting forward and backward features for best and second best models
            best_model_forward_features = final_features.get(best_model_name + "_forward", [])
            best_model_backward_features = final_features.get(best_model_name + "_backward", [])

        else:
            best_model_forward_features = embedded_features
            best_model_backward_features = embedded_features


        # Appending the results to the results dataframe
            new_result = {
            'Best Model': best_model_name,
            'Best Model Forward Features': best_model_forward_features,
            'Best Model Backward Features': best_model_backward_features,
            'Best Model Score': best_model_r2}

        results_df = results_df.append(new_result, ignore_index=True)

        # Appending to df in the .csv file
            pd.DataFrame([new_result]).to_csv(csv_file, mode = 'a', header=False, index=False)

        print(results_df)
            print('Data appended to', csv_file)
            print("-----------------------------------------------------------\n")
```

```
            count+=1

    return results_df

results_df = pipeline(df_num, 1000)
```

3. SOURCE CODE FOR FINAL ANALYSIS

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import re
import ast

# reading from the csv where iteration results were stored
df = pd.read_csv("./bootstrap_final.csv")

df.head()

df.info()

forward_list = []
# Converting string into a list
for i in df['Best Model Forward Features']:
            x = ast.literal_eval(i[7:-25])
            forward_list.append(x)
```

```
df['Best Model Forward Features'] = forward_list


backward_list = []
# Converting string into a list
for i in df['Best Model Backward Features']:
        x = ast.literal_eval(i[6:-23])
        backward_list.append(x)

df['Best Model Backward Features'] = backward_list


df.head()


total_forward_features = []
# Appending all forward features to a single list
for i in df['Best Model Forward Features']:
        total_forward_features+=i


len(total_forward_features)


total_backward_features = []
# Appending all backward features to a single list
for i in df['Best Model Backward Features']:
        total_backward_features+=i


# Appending all features to a single list
total_features = total_backward_features+total_forward_features
```

```
# Making a dictionary out of the list with distinct protein names and corresponding frequencies of them occuring in the list

frequency_dict = {item: total_features.count(item) for item in set(total_features)}


# Sorting the dictionary by their values (frequencies)

sorted_dict = {k: v for k, v in sorted(frequency_dict.items(), key=lambda item: item[1], reverse=True)}


len(sorted_dict)


# Unpacking keys and values

keys, values = zip(*sorted_dict.items())


# Plotting

plt.figure(figsize=(15, 10))

plt.barh(keys[:50], values[:50], color='skyblue')

plt.ylabel('Protein names')

plt.xlabel('Frequency')

plt.title('Frequency of the proteins')

plt.xticks(rotation=90)

plt.show()


# Reading the Cancer Cell Line dataset to get the protein and gene names to match with our analysis data

df_ccle = pd.read_excel("Table_S2_Protein_Quant_Normalized.xlsx", sheet_name = 1)


prot_gene_df = df_ccle[['Gene_Symbol','Description','Uniprot','Uniprot_Acc']]
```

```
# Making a df which is an intersection of all the genes relevant to breast cancer and the proteins found from our iterations

df_analysis = prot_gene_df[(prot_gene_df['Uniprot_Acc'].isin(keys))] # & (prot_gene_df['Gene_Symbol'].isin(bc_gene_list))



# Extract common items and their values from the dictionary

df_analysis['frequency'] = df_analysis['Uniprot_Acc'].map(sorted_dict)



df_analysis.head()



print(f'Number of unique values in ID: {df_analysis["Uniprot_Acc"].nunique()}')



# Sorting df by frequency

df_analysis = df_analysis.sort_values(by='frequency', ascending=False)

df_analysis.reset_index(inplace=True, drop=True)



df_analysis.head()



df_analysis = df_analysis[df_analysis['frequency']>150]



gene_list = list(df_analysis['Gene_Symbol'])



len(gene_list)



gene_df = pd.DataFrame(gene_list)



# Write the DataFrame to a CSV file

gene_df.to_csv('./final_gene.csv', index=False)
```

```
df_analysis.head()
```