MARMARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE KANSER İLE İLGİLİ YENİ BİYOBELİRTEÇLERİN TESPİT EDİLMESİ

FIRAT KURT

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği(İngilizce) Anabilim Dalı
Bilgisayar Mühendisliği Programı

DANIŞMAN
Doç.Dr. Mustafa AĞAOĞLU

EŞ-DANIŞMAN
Prof.Dr. Kazım Yalçın ARĞA

İSTANBUL, 2022

MARMARA UNIVERSITY
INSTITUTE FOR GRADUATE STUDIES
IN PURE AND APPLIED SCIENCES

IDENTIFICATION OF NOVEL SYSTEMS BIOMARKERS FOR CANCER DIAGNOSIS USING MACHINE LEARNING TECHNIQUES

FIRAT KURT

MASTER THESIS
Department of Computer Engineering

Thesis Supervisor
Assoc.Prof.Dr. Mustafa AĞAOĞLU

Thesis CO- Supervisor
Prof.Dr. Kazım Yalçın ARĞA

ISTANBUL, 2022

MARMARA ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Marmara Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Öğrencisi Fırat KURT'un "Identification of Novel Systems Biomarkers for Cancer Diagnosis using Machine Learning Techniques" başlıklı tez çalışması, 25 Temmuz 2022 tarihinde savunulmuş ve jüri üyeleri tarafından başarılı bulunmuştur.

Jüri Üyeleri

Doç.Dr. Mustafa AĞAOĞLU (Danışman)

Marmara Üniversitesi .....

Doç.Dr. Murat Can GANİZ (Üye)

Marmara Üniversitesi .....

Prof.Dr. Selim AKYOKUŞ (Üye)

Medipol Üniversitesi .....

ONAY

Marmara Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun … / … /2022 tarih ve 12345 sayılı kararı ile Fırat KURT'un Bilgisayar Mühendisliği (İngilizce) Anabilim Dalı Bilgisayar Mühendisliği Programında Yüksek Lisans derecesi alması onanmıştır.

Fen Bilimleri Enstitüsü Müdürü

Prof. Dr. Bülent EKİCİ

MARMARA UNIVERSITY
INSTITUTE FOR GRADUATE STUDIES IN PURE AND APPLIED SCIENCES

Fırat KURT, a Master of Science of Computer Engineering student of Marmara University Institute for Graduate Studies in Pure and Applied Sciences, defended his thesis entitled "Identification Of Novel Systems Biomarkers For Cancer Diagnosis Using Machine Learning Techniques", on July 25, 2022 and has been found to be satisfactory by the jury members.

Jury Members

Assoc.Prof.Dr. Mustafa AĞAOĞLU (Advisor)
Marmara University ....

Assoc.Prof.Dr Murat Can GANİZ (Jury Member)
Marmara University ....

Prof.Dr. Selim AKYOKUŞ (Jury Member)
Medipol University ....

APPROVAL

Marmara University Institute for Graduate Studies in Pure and Applied Sciences Executive Committee approves that Fırat KURT be granted the degree of Master of Science of Computer Engineering in Department of Computer Engineering, Computer Engineering Program on July …, 2022. (Resolution no: 12345).

Director of the Institute
Prof. Dr. Bülent EKİCİ

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my supervisor, Assoc.Prof.Dr. Mustafa Ağaoğlu, whose sincerity and encouragement I will never forget. Assoc.Prof.Dr. Ağaoğlu inspired me on the way to this master's degree. He is the true definition of a leader and the ultimate role model. This work would not have been possible without Prof.Dr. Kazım Yalçın Arğa, whose guidance enabled me to develop an understanding of bioinformatics from the first step of research. I am grateful for the extraordinary experiences he arranged for me and for giving me the opportunity to grow professionally. I am honored to be mentored by Assoc.Prof.Dr. Ağaoğlu and Prof.Dr. Arğa.

I am also grateful to my friends Abdurrehim Dağdelen, Mahmut Akarsu and Salih Eligüzel, for their extraordinary support. They did not leave me alone in my difficult times and did not withhold any support.

Finally, I would like to thank my wife Müjde Elif Kurt and my children Yusuf, Serra, and Semra. Their belief in me has kept my spirit and motivation up during this process. My heartfelt gratitude goes to my wife who always reminds me what is important in life and who always supports me during my research period.

TABLE OF CONTENTS

ÖZET

MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE KANSER İLE İLGİLİ YENİ BİYOBELİRTEÇLERİN TESPİT EDİLMESİ

Gelişmiş kanser tedavisi için hastalığın sınıflandırılması, hassas onkolojinin genel hedefine ulaşmak için kritik öneme sahiptir. Meme kanseri; risk faktörleri, etki, hastalık ilerleyişi, sağ kalım ve tedaviye yanıt verme açısından farklılık gösteren birden fazla moleküler etiyoloji ve alt tip ile dünya çapında en yaygın kötü tümörlerinden biridir. Bu nedenle, sadece hastalığın prognozunu değerlendirmek değil, aynı zamanda etkili ve bireyselleştirilmiş tedaviyi mümkün kılmak için meme kanserini moleküler alt tiplere ayırmada kesin stratejilerin geliştirilmesine açık bir ihtiyaç vardır. Burada, gen ekspresyonu seviyesindeki değişikliklerin sağlıklı bireylerin yanı sıra farklı meme kanseri alt tiplerine sahip hastaların farklılaşmasına yol açabileceğini varsaydık ve meme kanserinin moleküler alt tiplere doğru bir şekilde sınıflandırılmak için bir mRNA paneli geliştirmek üzere TCGA-BRCA (n=1.215) ve METABRIC (n=1.992) projelerinden kapsamlı klinik ve RNA-seq verilerini içeren bir makine öğrenimi yaklaşımı önerdik. Alternatif özellik eleme yöntemleri farklı sayıda özellik sayısı göz önünde bulundurularak özellik seçimi yapıldı ve akabinde farklı makine öğrenme yöntemleri bir araya getirilerek en iyi kombinasyonun sağlandığı birleştirilmiş öğrenme yöntemi kullandık ve sağlıklı bireylerin yanı sıra farklı meme kanseri alt tiplerine sahip hastaları sınıflandırmak için özellikler olarak 50 mRNA'yı içeren birleştirilmiş bir model belirledik. Modelin ayırt etme performansı, her bir alt türü sağlıklı topluluklardan ayırt eden ikili karşılaştırmalarda %100'e kadar ulaştı. Özellikler meme kanseri gelişiminde 2,25 ($p=5x10^{-7}$) risk oranı ile yüksek belirti performans gösterdi ve kanser gelişimi ile ilgili moleküler yolaklarla önemli ölçüde ilişkiliydi. Bu çalışma; topluluk öğrenmesi gibi ileri makine öğrenmesi yöntemleri uygulamalarının kanser hastalarının doğru sınıflandırmalarını ve hedefe yönelik hastalığın teşhis ve tedavi edilmesinde daha güçlü araçların geliştirilebilmesini sağlayabileceğini göstermektedir.

ABSTRACT

IDENTIFICATION OF NOVEL SYSTEMS BIOMARKERS FOR CANCER DIAGNOSIS USING MACHINE LEARNING TECHNIQUES

Stratification of disease for improved cancer treatment is critical to achieve the overall goal of precision oncology. Breast cancer is one of the most common malignancies worldwide, with multiple underlying molecular etiologies and subtypes that differ in risk factors, incidence, disease progression, survival, and response to treatment. Therefore, there is a clear need for the development of precise strategies to stratify breast cancer into molecular subtypes, not only to assess the prognosis of the disease, but also to enable effective and individualized therapy. Here, we hypothesized that changes at the level of gene expression could lead to differentiation of patients with different breast cancer subtypes as well as healthy individuals and proposed a machine learning approach incorporating comprehensive clinical and RNA-seq data from the TCGA-BRCA (n=1,215) and METABRIC (n=1,992) projects to develop an mRNA panel for accurate stratification of breast cancer into molecular subtypes. Using feature selection via alternative feature elimination methods, considering different feature sizes, and following an ensemble learning approach to find the best combination of various machine learning models, we identified an ensembled model including 50 mRNAs as features to stratify patients with different breast cancer subtypes as well as healthy individuals. The discrimination performance of the model reached up to 100% accuracy in pairwise comparisons to distinguish each subtype from the healthy cohort. The features showed high prognostic performance with hazard ratio of 2.25 ($p=5x10^{-7}$) in breast cancer development and were significantly associated with molecular pathways related to carcinogenesis. This study shows that we could consider the application of advanced machine learning methods, such as ensembled learning, for the accurate classification of cancer patients and provide more powerful tools for improving diagnosis and targeted therapy.

ABBREVIATIONS

AUC : Area Under Curve

CatBoost : Categorical Boosting

GDC : Genomics Data Commons

GO : Gene Ontology

FN : False Negative

FP : False Positive

HER2 : Human Epidermal Growth Factor Receptor 2

HR : Hazard Ratio

ICGC : International Cancer Genome Consortium

KEGG : Kyoto Encyclopedia of Genes and Genomes

K-M : Kaplan-Meier

KNN : k-Nearest Neighbor

LDA : Linear Discriminant Analysis

LightGBM : Light Gradient Boosting Machine

LumA : Luminal A

LumB : Luminal B

METABRIC : Molecular Taxonomy of Breast Cancer International Consortium

ML : Machine Learning

PCA : Principal Component Analysis

PI : Prognostic Index

RFE : Recursive Feature Elimination

ROC : Receiver Operating Characteristic

RPKM : Reads per kilo-base per million

SVM : Support Vector Machine

TCGA : The Cancer Genome Atlas

TN : True Negative

TNBC : Triple Negative Breast Cancer

TP : True Positive

XGBoost : Extreme Gradient Boosting

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

In recent decades, there has been an explosion of data. We live in an age of data, and there is a lot of structured and unstructured data. This data is useless unless it is analyzed to find the patterns hidden within it. Machine learning (ML) techniques automatically find the important information in complex data and turn this data into knowledge. The knowledge gained can be used for decision making. ML is an approach to learning complex patterns from existing data and using these patterns to make predictions about unseen data.

Traditionally, software development combines rules created by humans with data to find answers to a problem. Machine learning, on the other hand, uses data and answers to find the rules behind a problem (Chollet, 2017). ML provides the computer with the ability to learn without external programming (Samuel, 1959). ML is one of the main branches of Artificial Intelligence. ML is the intersection of statistics, mathematics, and computer science. The no-free-lunch theorem is famous in the field of machine learning. There is no single algorithm that can be used for all problems. Each problem has unique characteristics. Therefore, the most suitable ML algorithm must be found for each problem.

In the real world, raw data is usually not suitable for use in ML applications. Missing values, wrong types, differently scaled data or redundant information are the most common data problems. These problems must be eliminated before the data is used in ML algorithms. These processes are called data preprocessing and this step can be considered mandatory before using the data in ML algorithms. Data cleaning, feature selection, data scaling, feature engineering and dimensionality reduction are the most common preprocessing approaches.

ML algorithms extract information from existing data. Therefore, most ML algorithms require a training process. Training uses existing data to build a model. The ML algorithm learns parameter coefficients in the training process and applies these coefficients to unseen data to produce an output. In some cases, our model can outperform the training data. Therefore, the generated model shows excellent performance on the training data but poor performance on the test data. The performance of the generated model must be tested before using the model. The test data must not be used for training and must be split before the training process.

Each ML algorithm has its own approach to deal with problems. Therefore, the ensemble learning approach is used in state-of-the-art studies. In ensemble learning, the individual

algorithms of ML are used as an internal model and a meta-model is created from the output of these models. The actual output comes from the agreement of different ML models, *i.e.*, different approaches. The strength of ensemble learning results from this consensus.

ML plays an increasingly important role not only in computer science, but also in our everyday lives more than ever. We benefit from robust email spam filters, text and speech recognition applications, internet search engines, and chess-playing programs (Raschka and Mirjalili, 2019). Moreover, ML has a huge impact on various data-driven research fields. ML is used for various scientific problems such as understanding stars, searching for distant planets, analyzing DNA sequences, and delivering personalized cancer treatments.

The stratification of cancer patients and healthy individuals is crucial for early diagnosis and intervention in several cancers since the success of treatment strategies decreases as cancer progresses. Moreover, most cancers display molecular subtypes with different responses to conventional therapies. Therefore, the molecular characterization of patients into subtypes is also crucial for effective cancer treatment.

Biomarkers are needed both to distinguish early-stage cancer patients from diseased individuals and to characterize which cancer subtypes cancer patients belong to. Previous studies have shown that realistic and accurate classification based on a gene, mRNA, or protein alone is not possible, suggesting that a set of functionally related molecules (systems biomarkers) might be used for this purpose.

There are some behavioral differences between healthy cells and cancerous cells. These behavioral changes are caused by changes in the density of the proteins in the cell. These changes are determined by measuring the density of biomarkers in the cell. Biomarkers point to the genes responsible to produce these biomarkers. Therefore, gene sets that are altered in cancerous cells can be identified from changes in these biomarker densities. Biomarker density changes will be analyzed by machine learning methods and biomarkers varying density in cancerous cells will be determined. In this way, genes that are related to cancer will be identified with these biomarkers.

Although some success has been achieved with traditional statistical methods and developed bioinformatics tools in determining system biomarkers, biomarker sets with high specificity and high sensitivity are still not available for use in the clinic for certain cancers. Previous

studies also demonstrated that machine learning techniques may be more successful than traditional statistical methods in biomarker research.

Breast cancer is one of the most common malignancies worldwide and the second leading cause of cancer-related deaths in women. It is estimated that 287,850 new cases and 43,250 deaths from breast cancer are expected in the United States in 2022, and the probability of developing invasive cancer within the period from birth to death is estimated to be 12.9% (Siegel et al., 2022). In addition, the incidence rate of breast cancer in women has continued to increase by approximately 0.5% per year since the mid-2000s (Pfeiffer et al., 2018).

At the molecular level, gene expression profiling has transformed our understanding of breast cancer by identifying four predominant molecular subtypes: Luminal A (LumA), Luminal B (LumB), human epidermal growth factor receptor 2 (HER2)-enriched, and basal-like triple negative breast cancer (TNBC) (Johnson et al., 2021). The substantial differences between these molecular subtypes are a consequence of dramatically altered genomic and proteomic profiles that manifest as changes in activated signaling networks (Gatza et al., 2014) and are reflected in differences in risk factors, incidence, age, disease progression, survival, and response to treatment. Therefore, there is a clear need for the development of accurate strategies to stratify breast cancer into molecular subtypes, not only to assess the prognosis of the disease, but also to provide effective and individualized therapy.

Stratifying diseases for improved cancer treatment is critical to achieve the overall goal of personalized medicine. The promising role of machine learning approaches for the development and application of optimal medical practices in breast cancer has been demonstrated in several studies (Ozer et al., 2020). ML approaches have been applied to estimate breast cancer survival and predict outcomes with high accuracy (Montazeri et al., 2016; Turkki et al., 2019); however, little attention has been paid to molecular classification of breast cancer into molecular subtypes. Studies aimed at investigating the potential of ML for classifying breast cancer patients into molecular subtypes are very limited and focus on binary classification considering TNBC and non-TNBC subtypes (Wu and Hicks, 2021). On the other hand, publicly available population-wide genomic and transcriptomic profiling projects such as Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) (Curtis et al., 2012) and The Cancer Genome Atlas (TCGA) (Ciriello et al., 2015) provide sufficiently comprehensive molecular datasets to investigate subtype-specific molecular mechanisms of breast cancer.

Here, we hypothesized that changes at the level of gene expression could lead to the discrimination of patients with different breast cancer subtypes as well as healthy individuals and proposed an ML approach incorporating comprehensive clinical and RNA-seq data from METABRIC and TCGA projects to develop an mRNA panel for accurate stratification of breast cancer into molecular subtypes.

2. MATERIALS AND METHODS

2.1. Machine Learning Algorithms

2.1.1. Linear Discriminant Analysis (LDA)

Logistic regression is a simple and powerful linear classification algorithm. It has some limitations due to its simple structure. The main limitation is that it is appropriate for two-class classification. It may be used for multi-class classification, but it is not used for this purpose. Also, it is unstable for well-separated classes, and it is unstable for datasets that have few examples.

LDA addresses each of these limitations by the multivariate Gaussian method. In this method, the mean and the variance of the variable are calculated for each class. LDA makes some assumptions. These assumptions are dataset has Gaussian distribution and that each attribute of the dataset has the same variance. These assumptions simplify the algorithm, and the model estimates the mean and the variance from the dataset for each class (McLachlan, 2005).

LDA uses Bayes' Theorem for estimating the probability of each class. The base assumption is each data may belong to each class at the same probability. This is called the prior probability in Bayes' Theorem (Jeffreys, 1946).

P(Y=x|X=x) = (PIk * fk(x)) / sum(PIl * fl(x)), Plk= nk/n. Plk is the base probability of each class. The f(x) is estimated probability of x belonging to each class. Gaussian distribution rules are applied to f(x) function and each class classification probability is calculated by these new formulas. Dk(x) = x * (muk/siga^2) – (muk^2/(2*sigma^2)) + ln(PIk). Dk(x) is the discriminate function of class k for input x. Muk, sigma^2 and Plk are estimated from dataset.

2.1.2. Support Vector Machine (SVM)

SVM is one of the popular machine learning algorithms. It can be used in either classification or regression. SVM aims to find an N-dimensional hyperplane that classifies the data points (Platt, 1999). The number of features is assigned the dimension of the hyperplane. As can be seen in Figure 2.1(a), data can be classified by multiple lines. SVM finds a line that has the largest separation (margin) between two classes (Figure 2.1(b)). SVM does not always find the maximum margin. When the dataset contains an outlier (Figure 2.1(c)), SVM ignores these data points and finds the optimum margin between two classes (Figure 2.1(d)).

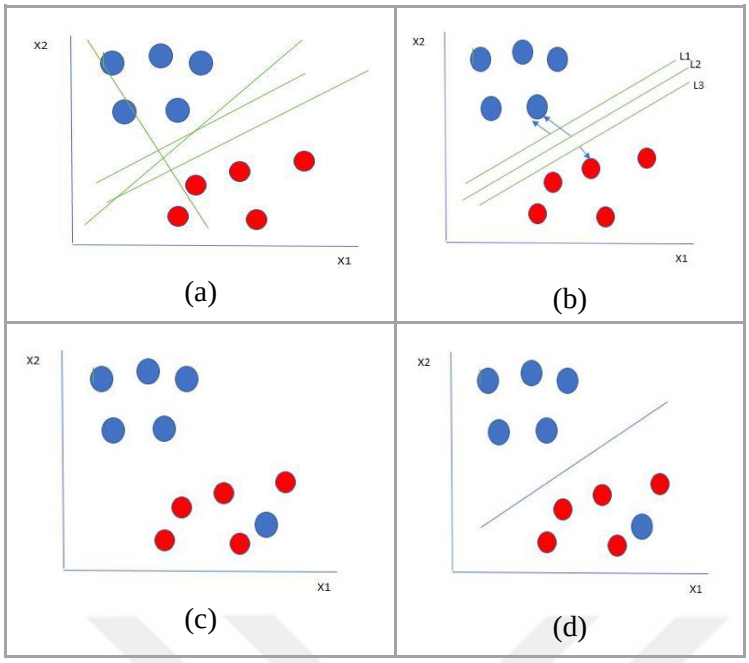


Figure 2.1. Support Vector Machine (SVM)

2.1.3. K-Nearest Neighbors (KNN)

KNN is a supervised learning algorithm which is used in classification and regression. KNN predicts the test data point's class by calculating the distance between all training data points and the test data. Then, we select K nearest train data points and calculate the probability of each class depending on these K training data points. The class which has the highest probability is selected. The KNN algorithm is a computationally expensive algorithm. This algorithm has no training time and distance between train data points and test data is calculated for all test data points one by one (Guo et al., 2004).

Choosing the correct K-value is an important part of this algorithm. Choosing too small or too big K-value affects the correctness of the model. To find the optimum K-value, a small K value is chosen, and the misclassification value is calculated for this K-value, then this value is increased one by one, and the misclassification value is calculated for each K-value. The K-value that minimizes the number of incorrect classifications is determined and this K-value is selected.

2.1.4. Ensemble Learning

Each machine learning method has its pros and cons. Ensemble learning is a meta-learning that generates better predicting ability by combining different machine learning models. Bagging, stacking and boosting are the main classes of ensemble learning.

- **Bagging**: Many decision trees are contained in bagging and these trees are trained by different sample sets which are generated from the same dataset. Prediction is done by averaging the results of these trees. Random Forest (Breiman, 2001) is the main sample of Bagging.
- **Stacking**: Many different model types are contained in staking and these models are trained by the same data. Results of these models are used in another model for predicting meta results (Wolpert, 1992).

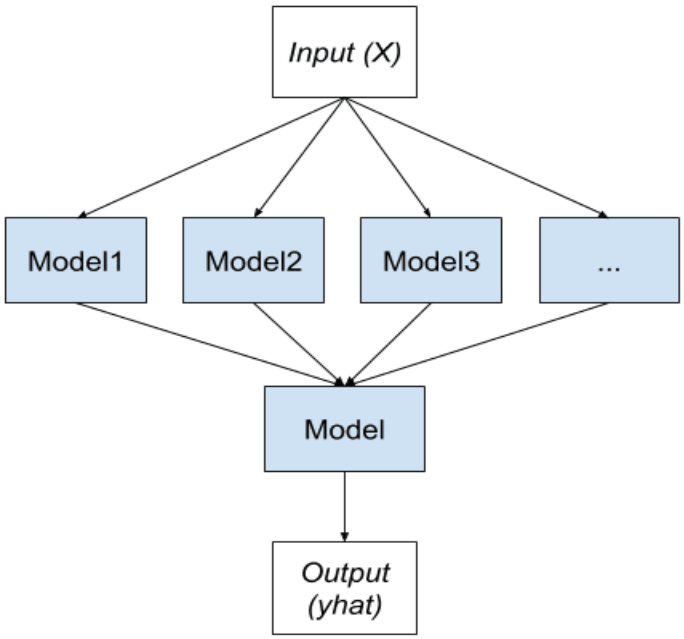


Figure 2.2 Stacking Ensemble

Stacking has its appellation for each level. Root level models are named as level-0. The resulting model of combining level-0 models is named as level-1. Two-level model is common, but we can add more levels. These extra levels are prone to overfitting, so we must be careful when adding more levels.

- **Boosting**: Ensemble members are added to the model sequentially in boosting. A new member is added for correcting previous model errors. Prediction is done by the weighted average of the results of these members.

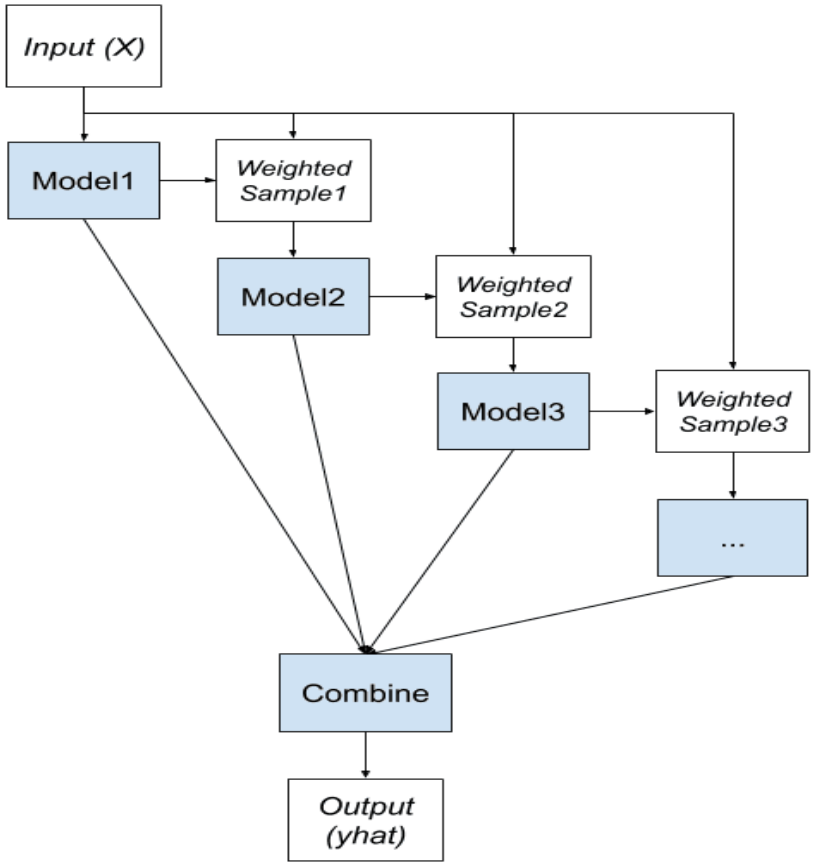


Figure 2.3 Boosting Ensemble

Correcting previous models' errors is the leading property of boosting ensembles. Each added model corrects the previous models' prediction errors. These models are added sequentially.

Simple decision trees (make single or a few decisions) are used in boosting and these trees are named weak learners. The predictions of weak learners are combined by voting or averaging approaches. But the contribution of models is weighted by their performance or capability. Boosting is a strong learner, and this strong learner is developed from many weak learners.

There are a lot of boosting models and the most popular three models were selected for this project. These models were LightGBM, XGBoost and CatBoost models, and they were subsamples of Gradient Boosting.

2.1.4.1. Gradient Boosting

Boosting is a general ensemble technique. Its structure is on the correction of the performance of previous models. Gradient Boosting improves the performance of the Boosting algorithm by taking some ideas from bootstrap aggregation, such as randomly sampling the samples and features when training the ensemble members. Gradient Boosting uses any lost function and gradient descent optimization algorithms.

There are three main reasons for performance improvement.

- Tree Constraints: The depth of the trees and the number of trees are arranged.
- Weighted Updates: The learning rate is used to limit the contribution of each tree to the ensemble.
- Random Sampling: Trees are trained random subsets of features and samples.

- **Light Gradient Boosted Machine (LightGBM)**

LightGBM has two main ideas: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). These two features accelerate the training speed up to 20x. GOSS is the modification of gradient descent. GOSS gives attention to the sample which has a large gradient. It speeds up the training process and reduces computational complexity. GOSS ignores much of the data which have a small gradient, and it obtains quite an accurate estimation with smaller data size. EFB is pruning mostly zero features such as categorical features (one-hot encoded). EFB can be thought of as an automatic feature selection. (Ke et al., 2017)

- **Extreme Gradient Boosting (XGBoost)**

Tianqi Chen who is the creator of XGBoost said that "It is designed to be both computationally efficient, fast in execution and extremely effective, perhaps even more effective than other open source implementations. However, the name XGBoost refers to the developers' goal of pushing the limits of computational resources for boosted tree algorithms. This is the reason why many people use XGBoost."

There are three main reasons for the performance improvement of XGBoost:

- XGBoost is one of the fastest versions of gradient boosting trees. The main problem of gradient boosting is finding optimum splitting points. XGBoost tackles this inefficiency of gradient boosting. XGBoost finds optimum splitting features by just looking at the distribution of features at the leaf to reduce the feature space of possible feature splits.
- XGBoost uses L1 and L2 regularization for tackling overfitting problems, so XGBoost has better generalization capability.
- One of the performance improvers of XGBoost is that XGBoost can work parallelized across clusters (Chen and Guestrin, 2016).

- **Categorical Boosting (CatBoost)**

CatBoost is a machine learning algorithm from Yandex. CatBoost name comes from "Category" and "Boosting" words. "Category" comes from CatBoost's ability to work with multiple data types, such as text, image, and audio. The main feature that distinguishes CatBoost from other boosting methods is that it can work with text data. "Boost" comes from gradient boosting machine learning algorithm. This algorithm is a gradient boosting algorithm.

CatBoost is powerful in two ways. It has good performance without extensive data training and supports many different data formats. Also, CatBoost reduces the need for hyperparameter tuning. Instead of other boosting methods, CatBoost is less prone to overfitting, so it has better generalization ability (Prokhorenkova, 2018).

2.2. Data Preprocessing

2.2.1. Feature Scaling

Many input datasets may not be used directly in machine learning models. Input datasets may have missing values, unnecessary features, different scaled features, etc. These datasets must be preprocessed before using in modeling. One of the most common forms of preprocessing is a simple linear scaling of the input variables (Bishop, 1995).

Modeling different scaled input dataset is very difficult. Large input values affect model parameters more than fewer feature values. In addition, these large parameters may change bigger than others and this effect model generalization performance is worst. These models are often unstable.

There are many feature scaling algorithms and normalization, standardization, and robust scaling are used in this study.

2.2.1.1. Normalization

Normalization is scaling input data between 0 and 1. This process does not change the distribution of the input data. $y = (x - min) / (max - min)$, where y is the normalized value, x is the original value, max is the maximum value of the input dataset and min is the minimum value of the input dataset.

2.2.1.2. Standardization

Standardization is rescaling the distribution of input values. This scaling changes the means of input data to 0 and the standard deviation of input data to 1. y = (x – mean) / stdev. The mean is calculated as mean = sum(x) / count(x), and the stdev is calculated as stdev = sqrt(sum((x – mean) ^2) / count(x)).

It is important that the initial mean and standard deviation must be calculated from just the training dataset, not the entire dataset. Otherwise, this scaling causes leakage and overfitting. The statistics required for the transformation are estimated from the training set and are applied to all datasets (Kuhn and Johnson, 2019).

This scaling can be thought of as centering the input data. Standardization assumes that input data has Gaussian distribution (bell curve). When the input data does not have Gaussian distribution, standardization may be used but it may not produce reliable results.

2.2.1.3. Robust Scaling

When the input dataset contains outliers, the mean and variance of the dataset are not calculated very well. These outliers are removed from the dataset in robust scaling. The statistic is used for scaling data in robust scaling. Data between some range (*i.e.*, 25^{th} and 75^{th} percentile range) are selected, and mean and variance are calculated depending on this selected data. It produces a more robust estimation for mean and variance.

2.2.2. Feature Selection

Feature selection is the process of reducing the dimension of the input dataset. The reduction of the input dataset either reduces the computational cost of modeling or (in some cases) improves the model's performance. Just as there is no best machine learning algorithm, there is no best feature selection method. The appropriate feature selection method must be discovered from careful experimentation with the input dataset. SelectKBest, RFE, and SelectFromModel feature selection algorithms, which are explained below, were used.

2.2.2.1. SelectKBest

Univariate Feature Selection is based on univariate statistical tests like chi2, Pearson correlation, etc. This statistical test is used for selecting features that have the strongest relationship with the output variable. SelectKBest is a class in the scikit-learn package, and it is combining univariate feature selection algorithms. F_classif, ANOVA F-value, chi2, and p-values can be used in the SelectKBest method (Pedregosa et al., 2011a).

2.2.2.2. Recursive Feature Elimination (RFE)

RFE is one of the popular feature selection algorithms. RFE is easy to configure and to use and it is effective at selecting features in the input dataset that are most relevant in predicting the target variables. RFE needs two important configurations: one is choosing the algorithm used to choose a feature, the other is the number of features that will be selected. However, the performance of RFE is not strongly dependent on these hyper-parameters (Guyom et al., 2002).

RFE uses different machine learning algorithms in the core for helping in features selection, so it is called wrapper-type feature selection. RFE starts working with all features in the input dataset. The machine learning algorithm used in the core of the model ranks features by importance and discards the least important features. This process is repeated until a specified number of features remains.

2.2.2.3. SelectFromModel

SelectFromModel is similar to RFE. The main difference is that SelectFromModel is based on the threshold of the importance attribute (coef_ or feature_importances_). Therefore, the machine learning algorithm used in the core of the model must have the coef_ or feature_importances_ attributes. The feature importance is the value of the influence of the input features on the prediction. The importance values are calculated for all features. The features are sorted by their importance values and the most valuable features are selected depending on the desired number of features. (Ferri et al., 1994).

2.3. Gene Expression Datasets

Clinical and RNA-seq data from breast cancer patients and healthy individuals were obtained from two population-wide genomic and transcriptomic profiling projects, The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013) and the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) (Curtis et al., 2012).

2.3.1. Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)

METABRIC datasets containing RPKM-normalized expression data (EGAD00010000210, EGAD00010000211, and EGAD00010000212, n=1,992) for 25,186 probes were obtained from the European Genome-Phenome Archive (https://ega-archive.org) and employed under the permission from the ICGC Data Access Committee.

2.3.2. The Cancer Genome Atlas (TCGA)

The TCGA-BRCA dataset associated with breast invasive carcinoma (TCGA-BRCA, n=1,215) was downloaded from the National Cancer Institute Genomics Data Commons (GDC) data portal (https://portal.gdc.cancer.gov/projects/TCGA-BRCA) and included reads per kilo-base per million mapped (RPKM)-normalized gene expression data for 56,908 probes.

METABRIC datasets were used for model development and training and testing purposes, while the TCGA-BRCA dataset was used for external validation of results and in biological interpretation of the results through survival analysis.

2.4. Implementation

2.4.1. Data Preprocessing

Most machine learning algorithms require some preparation of the dataset. This is called preprocessing. There are many different preprocessing operations: data cleaning, train-test splitting, feature selection, and feature scaling are applied to the dataset.

2.4.1.1. Data Cleaning

Either TCGA-BRCA or METABRIC datasets contain some records that contain redundant information. 225 TCGA-BRCA records do not contain subcategory information. In addition, the Normal subcategory is not included in the METABRIC dataset. Therefore, these two types of records are removed from the TCGA-BRCA dataset. On the other hand, the METABRIC dataset contains 6 records in the subcategory “NC”. This subcategory is not included in the TCGA-BRCA dataset. These records are removed from the METABRIC dataset.

2.4.1.2. Train Test Splitting

In this study, a scaled METABRIC dataset is used, and no information on scaling methods is available. Therefore, the TCGA-BRCA and METABRIC datasets cannot be scaled together, and one dataset cannot be used for training and the other for testing. Both datasets are divided into training and testing datasets. The METABRIC dataset has more data, so an 80/20 distribution was chosen for the METABRIC dataset. The TCGA-BRCA dataset has fewer data sets, therefore distribution of almost 90/10 was chosen for the TCGA-BRCA dataset.

Table 2.1 TCGA-BRCA and METABRIC Data Distribution

Subcategory	TCGA-BRCA			METABRIC		
	Train	**Test**	**Total**	**Train**	**Test**	**Total**
Basal	129	10	139	265	66	331
Her2	57	11	68	192	48	240
LumA	431	38	469	576	145	721
LumB	149	19	168	393	99	492
Healthy	105	8	113	162	40	202
TOTAL	**871**	**86**	**957**	**1,588**	**398**	**1,986**

2.4.1.2. Feature Selection

The main goal of this study is to find new biomarkers for breast cancer subtype identification. Both datasets contain a lot of biomarker information, and the most relevant ones for breast cancer need to be identified. Feature selection algorithms are used to find the most relevant biomarkers. There are many different feature selection algorithms. For this study, the SelectFromModel, RFE, and SelectKBest algorithms were selected.

There is no information on how many features need to be selected. Selecting too many biomarkers may be aimless, and selecting fewer features may result in underfitting. To find the optimal number of biomarkers, different numbers (20, 25, 30, 50, 100) of biomarkers were selected. At the end, 15 datasets are created for each dataset.

Table 2.2 Feature Selection Datasets Created

#	Dataset
1	SelectFromModel_20
2	SelectFromModel_25
3	SelectFromModel_30
4	SelectFromModel_50
5	SelectFromModel_100

#	Dataset
6	RFE_20
7	RFE_25
8	RFE_30
9	RFE_50
10	RFE_100

#	Dataset
11	SelectKBest_20
12	SelectKBest_25
13	SelectKBest_30
14	SelectKBest_50
15	SelectKBest_100

2.4.1.3. Feature Scaling

Feature scaling affects the performance of some machine learning algorithms. There is no optimal feature scaling algorithm for all machine learning algorithms. MinMaxScaling, StandardScaling, and RobustScaling are the most common feature scaling methods and are therefore selected.

The TCGA-BRCA dataset does not contain scaled data, so these feature scaling methods were applied to the TCGA-BRCA dataset, but the METABRIC dataset is also scaled, so the METABRIC dataset is not rescaled. There are 45 different datasets generated for the TCGA-BRCA dataset and 15 different datasets generated for the METABRIC dataset.

Table 2.3 All TCGA-BRCA Datasets Created

#	Dataset	#	Dataset	#	Dataset
1	SelectFromModel_100-MinMaxScaler	16	RFE_100-MinMaxScaler	31	SelectKBest_100-MinMaxScaler
2	SelectFromModel_100-StandardScaler	17	RFE_100-StandardScaler	32	SelectKBest_100-StandardScaler
3	SelectFromModel_100-RobustScaler	18	RFE_100-RobustScaler	33	SelectKBest_100-RobustScaler
4	SelectFromModel_20-MinMaxScaler	19	RFE_20-MinMaxScaler	34	SelectKBest_20-MinMaxScaler
5	SelectFromModel_20-StandardScaler	20	RFE_20-StandardScaler	35	SelectKBest_20-StandardScaler
6	SelectFromModel_20-RobustScaler	21	RFE_20-RobustScaler	36	SelectKBest_20-RobustScaler
7	SelectFromModel_25-MinMaxScaler	22	RFE_25-MinMaxScaler	37	SelectKBest_25-MinMaxScaler
8	SelectFromModel_25-StandardScaler	23	RFE_25-StandardScaler	38	SelectKBest_25-StandardScaler
9	SelectFromModel_25-RobustScaler	24	RFE_25-RobustScaler	39	SelectKBest_25-RobustScaler
10	SelectFromModel_30-MinMaxScaler	25	RFE_30-MinMaxScaler	40	SelectKBest_30-MinMaxScaler
11	SelectFromModel_30-StandardScaler	26	RFE_30-StandardScaler	41	SelectKBest_30-StandardScaler
12	SelectFromModel_30-RobustScaler	27	RFE_30-RobustScaler	42	SelectKBest_30-RobustScaler
13	SelectFromModel_50-MinMaxScaler	28	RFE_50-MinMaxScaler	43	SelectKBest_50-MinMaxScaler
14	SelectFromModel_50-StandardScaler	29	RFE_50-StandardScaler	44	SelectKBest_50-StandardScaler
15	SelectFromModel_50-RobustScaler	30	RFE_50-RobustScaler	45	SelectKBest_50-RobustScaler

Table 2.4 All METABRIC Datasets Created

#	Dataset
1	SelectFromModel_20
2	SelectFromModel_25
3	SelectFromModel_30
4	SelectFromModel_50
5	SelectFromModel_100

#	Dataset
6	RFE_20
7	RFE_25
8	RFE_30
9	RFE_50
10	RFE_100

#	Dataset
11	SelectKBest_20
12	SelectKBest_25
13	SelectKBest_30
14	SelectKBest_50
15	SelectKBest_100

2.4.2. Machine Learning Models Generation

2.4.2.1. Machine Learning Models Selection

There are many different machine learning algorithms. The performance of the machine learning algorithm depends on the input data set. Therefore, machine learning models must be trained with the input dataset to determine the machine learning algorithms with the best results. LazyPredict is a Python package that computes the performance of 29 different machine learning algorithms (Pandala, 2022). The performance of the ML algorithms is listed below.

Table 2.5 Performance Results of LazyPredict ML Algorithms

Model	Accuracy	Balanced Accuracy	F1 Score
NearestCentroid	0.77	0.81	0.79
LinearDiscriminantAnalysis	0.85	0.80	0.85
GaussianNB	0.72	0.78	0.74
BernoulliNB	0.79	0.77	0.80
KNeighborsClassifier	0.83	0.75	0.82
LogisticRegression	0.85	0.74	0.85
QuadraticDiscriminantAnalysis	0.82	0.74	0.82
LGBMClassifier	0.85	0.72	0.84
XGBClassifier	0.84	0.72	0.84
SVM	0.84	0.71	0.83
CatBoostClassifier *	0.81	0.71	0.81

* CatBoostClassifier is not contained in LazyPredict so the performance of CatBoostClassifier is calculated without the LazyPredict package.

2.4.2.2. Ensembling The Selected Machine Learning Models

Ensemble learning is meta-learning that produces better predictive capabilities by combining different machine learning algorithms. It can be seen that many models in the model performance report have similar performance values. Considering the performance and infrastructure of the models; XGBoost, LGBM, CatBoost, SVC, KNN and LDA were selected.

An ensemble model is created from these 6 machine learning algorithms. A two-level blending model was created. In Level-0, meta models were created by applying a 5-in-6 combination to selected ML algorithms. The results of the Level-0 meta models are generated by combining the results of these 5 ML models. The VotingClassifier is used to generate the meta-model results (Pedregosa et al., 2011b).

Table 2.6 Level-0 Blending Models

#	**Selected ML 1**	**Selected ML 2**	**Selected ML 3**	**Selected ML 4**	**Selected ML 5**
1	XGBoost	LGBM	CatBoost	SVC	KNN
2	XGBoost	LGBM	CatBoost	SVC	LDA
3	XGBoost	LGBM	CatBoost	KNN	LDA
4	XGBoost	LGBM	SVC	KNN	LDA
5	XGBoost	CatBoost	SVC	KNN	LDA
6	LGBM	CatBoost	SVC	KNN	LDA

Level-1 meta-model is generated from Level-0 meta-model results. RandomForestClassifier is used as a blender model. RandomForestClassifier takes results of these six models as an input dataset and produces the result of the ensembled model.

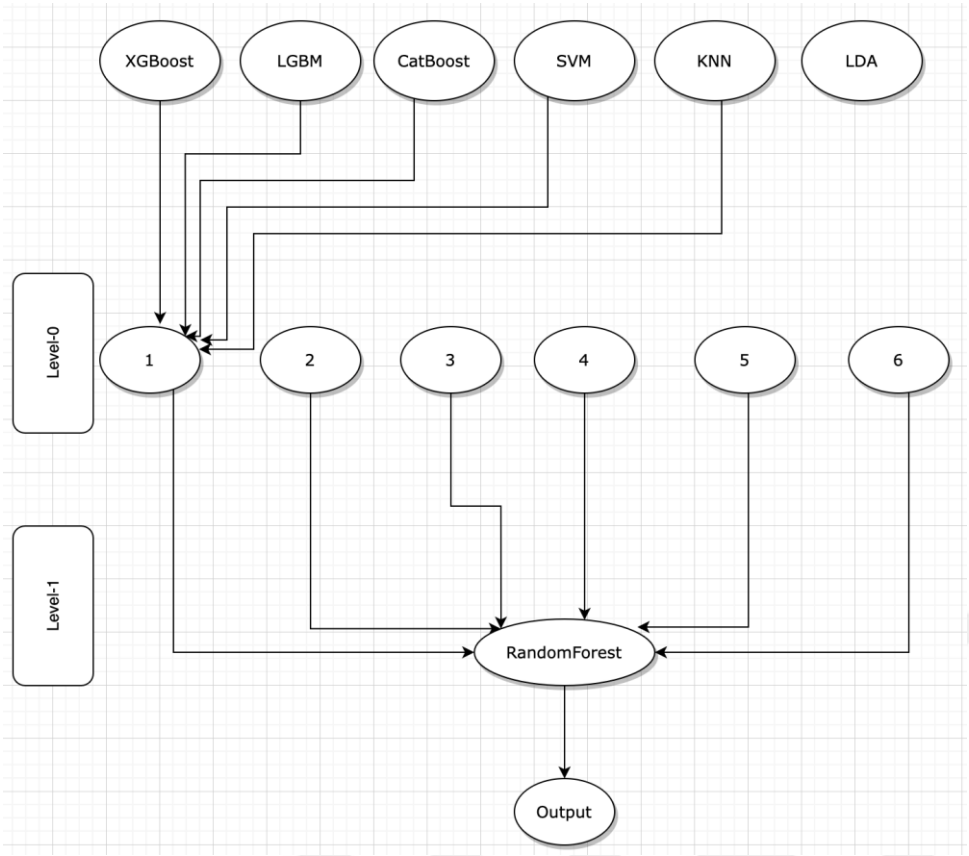


Figure 2.4. Ensemble Model Created

2.4.2.3. Hyperparameter Tuning

Some parameters of the ML algorithms can be obtained as a result of the training process. These parameters (K value for KNN, C value for SVM, etc.) have very important effects on the performance of the models. KNN, SVM, LGBM and XGBoost require tuning of hyperparameters. Custom hyperparameter settings are written for KNN and SVM, and Optuna (Akiba, 2019) is used for hyperparameter setting for LGBM and XGBoost. Optuna is a Python package that finds optimal parameter sets from given possible parameter sets.

Moreover, these parameter values may change depending on the input data set. Therefore, these parameters need to be calculated for each input dataset. Therefore, the hyperparameters of these ML algorithms are calculated for each input data set at the beginning of the model training.

2.4.3. Performance Metrics

2.4.3.1 Accuracy, Precision, and Recall

45 TCGA-BRCA datasets and 15 METABRIC datasets were created. These datasets were created to detect biomarkers highly associated with breast cancer. The ML model created was sequentially trained on these datasets and the performance of these datasets was measured. These main datasets (TCGA-BRCA and METABRIC) were not normally distributed datasets, so the accuracy value was not sufficient for measurement. Therefore, the precision and recall values were calculated to compare the performances of the datasets.

2.4.3.2 Confusion Matrix, ROC Curve, and PCA

A confusion matrix was constructed based on the results of each simulation to determine true positive (TP), true negative (TN), false positive (FP) and false negative (FN) metrics. The classification results of the models were evaluated by calculating the overall accuracy, precision, and recall indicators as well as the Area Under Curve (AUC) metric of Receiver Operating Characteristic (ROC) curve analysis. In addition, Principal Component Analysis (PCA) was performed to visualize the classification performance.

2.4.3.3 Gene Set Over-representation Analysis

Over-representation analysis were carried out using ConsensusPathDB (ver.35) (Kamburov et al., 2013) to recognize functional annotations such as metabolic, signaling and disease pathways, and biological processes significantly related to the gene products. Preferred data sources for metabolic and signaling pathways was Kyoto Encyclopedia of Genes and Genomes (KEGG, release 102.0) (Kanehisa et al., 2017), and Gene Ontology (GO) annotations (Ashburner et al., 2000) were employed to represent associated biological processes. p-values were obtained to determine significant biological processes and pathways through Fisher's exact test followed by Benjamini-Hochberg's correction as the multiple testing correction method. The enrichment results having adjusted $p<0.05$ were considered statistically significant.

2.4.3.4 Survival Analysis

To determine the association of selected feature set (mRNAs) with patients' overall survival, we performed a survival analysis using clinical data from the TCGA-BRCA cohort. The survival analysis was performed on 930 patients with available clinical information. Survival analysis was performed based on individual mRNA expression using the Kaplan-Meier (K-M)

approach through the survival package in R, by dividing the subjects into two groups (high- and low-risk) according to their prognostic index (PI), which is the linear component of the Cox proportional hazard model (PI = $\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_N x_N$, where β_i is the coefficient obtained from the Cox fitting and x_i is the expression value of each gene). The signatures of survival in each group were estimated by K-M plots. Plot with log-rank p-value<0.05 was considered as statistically significant and hazard ratio (HR = (O1/E1)/(O2/E2)) was calculated to discover the significance of survival plot based on the ratio between the relative death rate in group 1 (O1/E1) and group 2 (O2/E2), where O represents the observed number of deaths and E represents the expected number of deaths.

3. RESULTS AND DISCUSSION

3.1. Model Construction

To identify a set of mRNAs that could lead to discrimination of patients with different breast cancer subtypes as well as healthy individuals, three alternative feature selection algorithms (RFE, SelectKBest, and SelectFromModel) are commonly used to eliminate features of low importance from a training dataset were applied to the METABRIC training cohort (Table 3.1) and to the TCGA-BRCA training cohort (Table 3.1).

Table 3.1. The Distribution of Samples into Breast Cancer Subtypes

Subtype	**METABRIC Internal Train**	**METABRIC Internal Validation**	**TCGA-BRCA External Train**	**TCGA-BRCA External Validation**
Normal (Healthy)	162	40	105	8
Basal (TNBC)	265	66	129	10
Her2+	192	48	57	11
LumA	576	145	431	38
LumB	393	99	149	19
TOTAL	**1,588**	**398**	**871**	**86**

The models with which each dataset works in harmony are different due to the nature of the dataset and disadvantages of the modeling algorithms. Therefore, we considered the combined use of algorithms and followed an ensembled learning approach to obtain the best combination of models by using five out of the six algorithms in each simulation. Considering feature sets of five different sizes (*i.e.*, 20, 25, 30, 50, and 100), as a result, 15 different mRNA sets were created.

3.2. Performance Evaluations

The performance of the ensembled models was evaluated on the datasets test cohort as an internal validation to select the best model (*i.e.*, the most appropriate feature set) in stratification of patients into molecular subtypes. In addition to an overall analysis to stratify all subtypes and the healthy cohort, pairwise subtype analyzes to distinguish each subtype from the healthy cohort were also considered when evaluating the performance of the models. In addition to the indicators of accuracy, precision, and recall derived from the confusion matrices, the AUC metric of the ROC curve analysis and PCA visualizations were also considered in the evaluation of model performance.

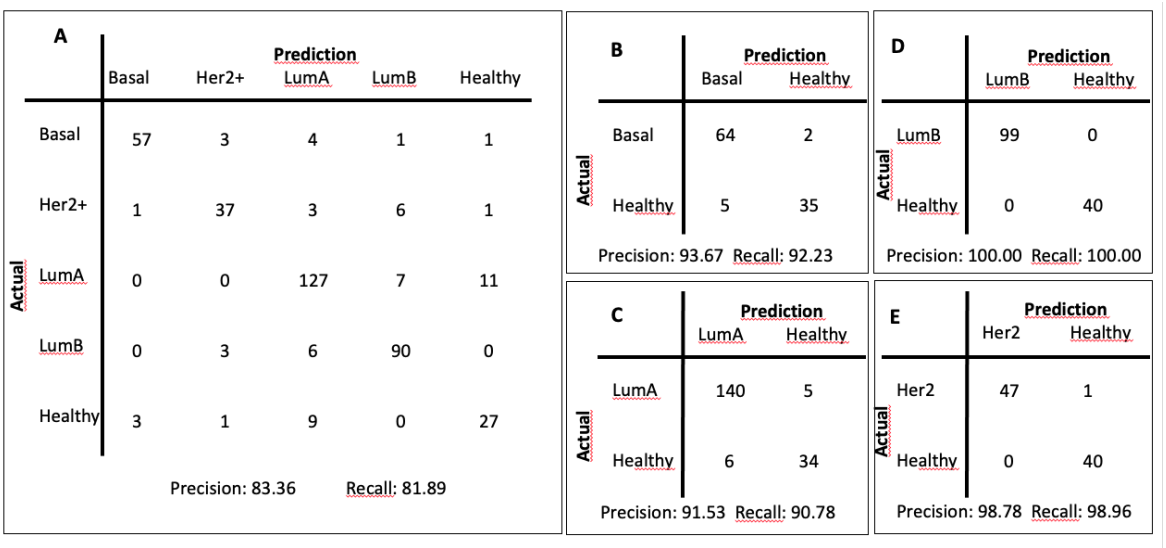
Performance evaluations on the datasets test cohort based on overall accuracy, precision, and recall in stratifying patients into molecular subtypes underscored several models (Table 3.2 and Table 3.3). Further analyzes based on the confusion matrices (Figure 3.1), AUC metrics of ROC -curve analyzes (Figure 3.2), and PCA visualizations (Figure 3.3 and 3.4) indicated the high performance of the feature set (so-called RFE-50) with a set size of 50 mRNAs selected by RFE.

Table 3.2. Model Performance Results on METABRIC Test Cohort

Feature Set Size	Feature Elimination Method	Accuracy	Precision	Recall
20	RFE	0.8216	0.8252	0.7773
	SelectKBest	0.7588	0.7646	0.7200
	SelectFromModel	0.7412	0.7200	0.6768
25	RFE	0.8568	0.8737	0.8183
	SelectKBest	0.7437	0.7805	0.6977
	SelectFromModel	0.7739	0.7471	0.7236
30	RFE	0.8191	0.7959	0.8020
	SelectKBest	0.8090	0.7907	0.7643
	SelectFromModel	0.8366	0.8229	0.8152
50	RFE	0.8291	0.8402	0.8020
	SelectKBest	0.7965	0.7925	0.7536
	SelectFromModel	0.8090	0.8017	0.7728
100	RFE	0.8392	0.8422	0.8076
	SelectKBest	0.7990	0.7805	0.7511
	SelectFromModel	0.7965	0.7803	0.7780

Table 3.3. Model Performance Results on TCGA-BRCA Test Cohort

Feature Set Size	Feature Elimination Method	Feature Elimination Method	Accuracy	Precision	Recall
20	FeatureSelection	MinMaxScaler	0.7889	0.6773	0.6938
		StandardScaler	0.8111	0.6847	0.7026
		RobustScaler	0.8111	0.8739	0.7683
	RFE	MinMaxScaler	0.7444	0.7067	0.6501
		StandardScaler	0.7556	0.7223	0.6861
		RobustScaler	0.7667	0.7240	0.6885
	SelectKBest	MinMaxScaler	0.7000	0.6324	0.5887
		StandardScaler	0.6556	0.6636	0.5560
		RobustScaler	0.6444	0.6057	0.5492
25	FeatureSelection	MinMaxScaler	0.7222	0.5878	0.6014
		StandardScaler	0.7444	0.6158	0.6317
		RobustScaler	0.7222	0.6048	0.6014
	RFE	MinMaxScaler	0.7556	0.6955	0.6861
		StandardScaler	0.7444	0.6868	0.6817
		RobustScaler	0.7778	0.7165	0.6992
	SelectKBest	MinMaxScaler	0.6778	0.6115	0.5711
		StandardScaler	0.6778	0.6315	0.5652
		RobustScaler	0.7222	0.7133	0.5880
30	FeatureSelection	MinMaxScaler	0.7889	0.8547	0.7267
		StandardScaler	0.7556	0.6669	0.6547
		RobustScaler	0.7889	0.8574	0.7331
	RFE	MinMaxScaler	0.7222	0.6831	0.6286
		StandardScaler	0.7556	0.7232	0.6676
		RobustScaler	0.7333	0.7029	0.6874
	SelectKBest	MinMaxScaler	0.6889	0.6088	0.5691
		StandardScaler	0.6222	0.6556	0.5046
		RobustScaler	0.6889	0.6731	0.5823
50	FeatureSelection	MinMaxScaler	0.7889	0.8599	0.7287
		StandardScaler	0.8000	0.8623	0.7331
		RobustScaler	0.8233	0.8818	0.7506
	RFE	MinMaxScaler	0.7667	0.6274	0.6655
		StandardScaler	0.7444	0.7008	0.6680
		RobustScaler	0.7333	0.7143	0.6994
	SelectKBest	MinMaxScaler	0.7444	0.6335	0.6631
		StandardScaler	0.6889	0.5815	0.6304
		RobustScaler	0.7333	0.6350	0.6631
100	FeatureSelection	MinMaxScaler	0.7889	0.7683	0.7208
		StandardScaler	0.8444	0.8513	0.7995
		RobustScaler	0.8222	0.8694	0.7506
	RFE	MinMaxScaler	0.7778	0.7289	0.7159
		StandardScaler	0.7667	0.7454	0.6964
		RobustScaler	0.7889	0.7710	0.7139
	SelectKBest	MinMaxScaler	0.7889	0.6906	0.6894
		StandardScaler	0.8000	0.7060	0.6982
		RobustScaler	0.7778	0.6677	0.6850



A

Actual \ Prediction	Basal	Her2+	LumA	LumB	Healthy
Basal	57	3	4	1	1
Her2+	1	37	3	6	1
LumA	0	0	127	7	11
LumB	0	3	6	90	0
Healthy	3	1	9	0	27

Precision: 83.36 Recall: 81.89

B

Actual \ Prediction	Basal	Healthy
Basal	64	2
Healthy	5	35

Precision: 93.67 Recall: 92.23

C

Actual \ Prediction	LumA	Healthy
LumA	140	5
Healthy	6	34

Precision: 91.53 Recall: 90.78

D

Actual \ Prediction	LumB	Healthy
LumB	99	0
Healthy	0	40

Precision: 100.00 Recall: 100.00

E

Actual \ Prediction	Her2	Healthy
Her2	47	1
Healthy	0	40

Precision: 98.78 Recall: 98.96

Figure 3.1. RFE_50 Confusion Matrices

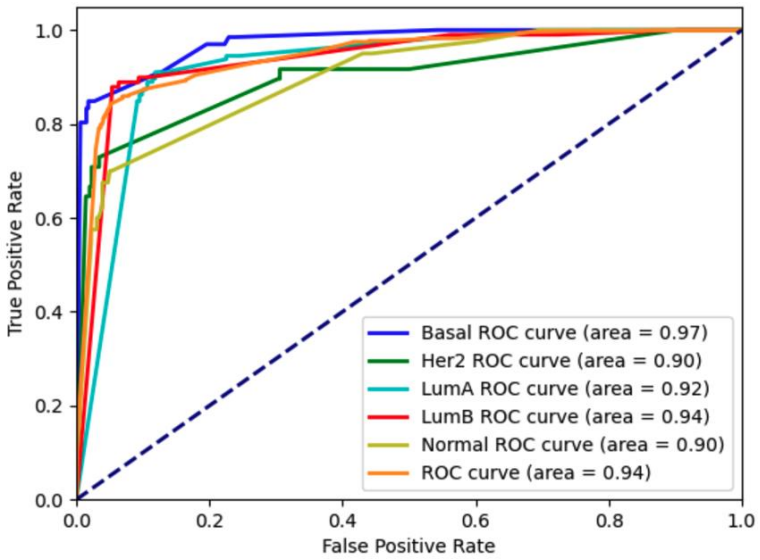


Figure 3.2. RFE_50 ROC Curve

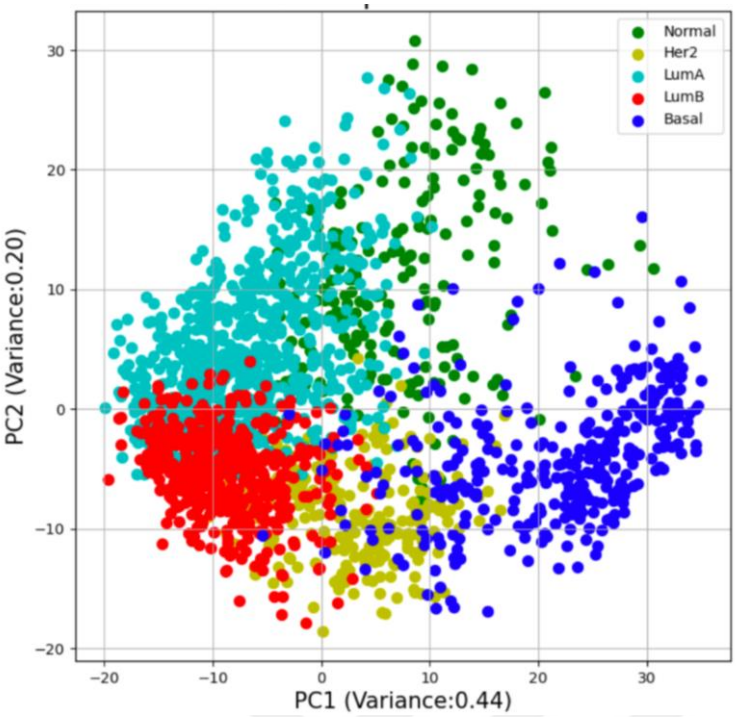


Figure 3.3. RFE_50 PCA Analysis for All Subtypes

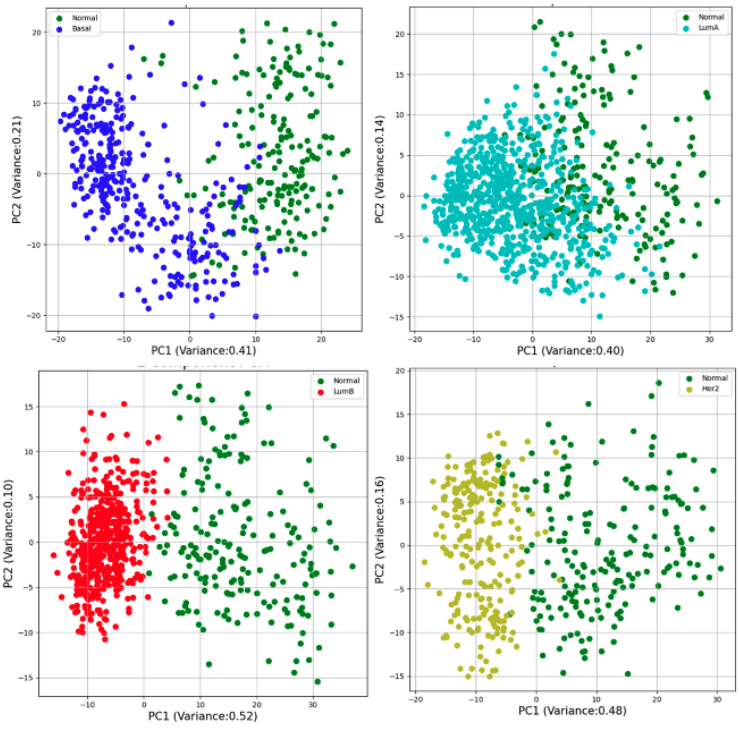


Figure 3.4. RFE_50 PCA Analysis for Normal (Healthy) vs Other

The RFE_50 model achieved 84.92% accuracy, 83.36% precision, 81.89% recall and an AUC value of 0.94 in stratifying patients into molecular subtypes (Figures 3.1 and 3.2). However, in pairwise comparisons to distinguish each subtype from the healthy cohort, the discrimination performance was significantly higher, e.g., 100% accuracy in distinguishing patients with LumB from healthy individuals and 98.86% accuracy in distinguishing patients with Her2+ from healthy individuals (Figure 3.1).

3.3. Ensemble Model Performance Comparison with Individual ML models

6 ML algorithms were selected to create an ensemble learning model. The main idea in creating the model was to increase the values for accuracy, precision, and recall. Comparisons were made between the "Normal" and "Breast Cancer" subtypes for all models of ML, as well as performance values for the entire type. The SVM algorithm performed better than the other models when comparing "Normal" and "Basal" subtypes (Table 3.4). Our model performed better than the other models when comparing "Normal" and "LumA" subtype (Table 3.5). SVM, LightGBM, and our model provided 100% separation when comparing "Normal" and "LumB" subtype (Table 3.6). This is an expected result, as the RFE_50 dataset was selected based on its success in separating "Normal" and "LumB" subtype. LDA outperformed the other models when comparing "Normal" and "Her2" subtype (Table 3.7). When comparing "Normal" and "All Breast Cancer" subtypes (Table 3.8) simultaneously, our model performed better than the others.

SVM and LDA are successful algorithms for two-class classification. Therefore, they achieved the best results in some comparisons. However, our model outperformed all models in both two-class and multi-class classification. For multiclass classification, our model outperformed the other models by at least two percent.

Table 3.4. Normal vs Basal Subtype

ML Model	Accuracy	Precision	Recall
LDA	0.9057	0.8973	0.9045
SVM	**0.9717**	**0.9723**	**0.9674**
KNN	0.9057	0.8996	0.8996
LightGBM	0.9623	0.9648	0.9549
XGBoost	0.9528	0.9520	0.9473
CatBoost	0.9623	0.9598	0.9598
Our Model	0.9340	0.9367	0.9223

Table 3.5. Normal vs LumA Subtype

ML Model	Accuracy	Precision	Recall
LDA	0.9027	0.8794	0.8203
SVM	0.8919	0.8442	0.8315
KNN	0.8865	0.8341	0.8280
LightGBM	0.8973	0.8504	0.8440
XGBoost	0.8432	0.7719	0.7461
CatBoost	0.9081	0.8933	0.8237
Our Model	**0.9405**	**0.9153**	**0.9078**

Table 3.6. Normal vs LumB Subtype

ML Model	Accuracy	Precision	Recall
LDA	0.9784	0.9853	0.9625
SVM	**1.0000**	**1.0000**	**1.0000**
KNN	0.9784	0.9853	0.9625
LightGBM	**1.0000**	**1.0000**	**1.0000**
XGBoost	0.9784	0.9853	0.9625
CatBoost	0.9928	0.9950	0.9875
Our Model	**1.0000**	**1.0000**	**1.0000**

Table 3.7. Normal vs Her2 Subtype

ML Model	Accuracy	Precision	Recall
LDA	**0.9886**	**0.9878**	**0.9896**
SVM	0.9318	0.9318	0.9354
KNN	0.9659	0.9668	0.9646
LightGBM	0.9205	0.9193	0.9208
XGBoost	0.9091	0.9105	0.9062
CatBoost	0.9659	0.9668	0.9646
Our Model	**0.9886**	**0.9878**	**0.9896**

Table 3.8. Normal and All Subtypes

ML Model	Accuracy	Precision	Recall
LDA	0.7864	0.7748	0.7604
SVM	0.7915	0.7768	0.7843
KNN	0.7965	0.7806	0.7588
LightGBM	0.8291	0.8289	0.7989
XGBoost	0.7990	0.7779	0.7549
CatBoost	0.8266	0.8125	0.7837
Our Model	**0.8492**	**0.8336**	**0.8189**

3.4. External Validation

For The RFE_50 model, further performance evaluations were performed on TCGA-BRCA cohort as external validation purposes. Our model represented 82.55% accuracy, 84.10% precision, 84.55% recall in stratifying patients into molecular subtypes; however, its discrimination performance was perfect with 100% accuracy in distinguishing Basal, LumB, and Her2+ subtypes from healthy individuals, and 95.65% accuracy in distinguishing patients with LumA from healthy individuals (Figure 3.5, Figure 3.6 and Figure 3.7).

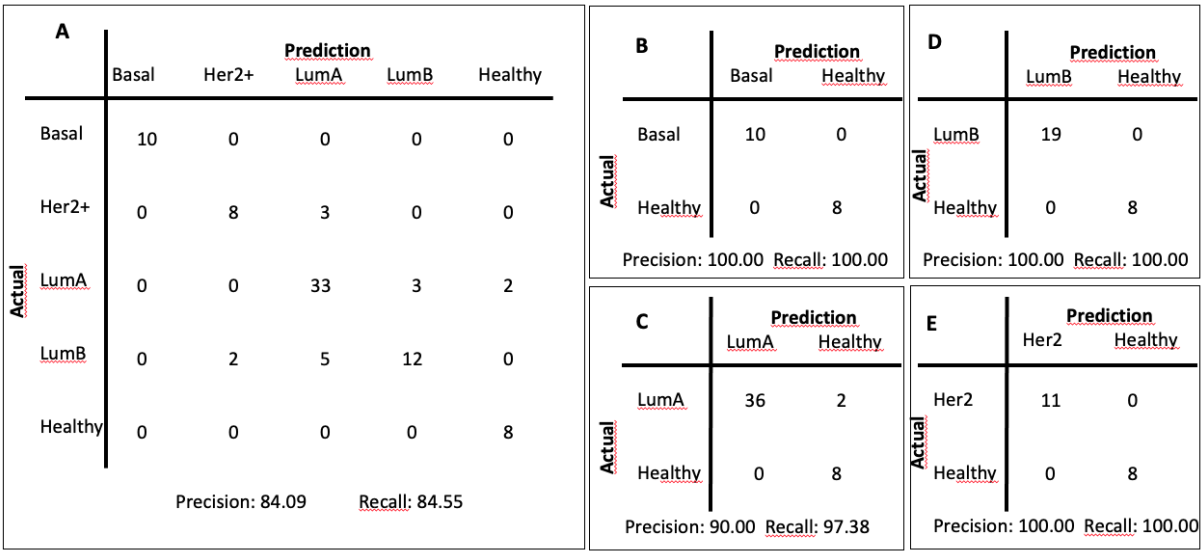


Figure 3.5. RFE_50 Feature Set Performance on TCGA-BRCA Dataset

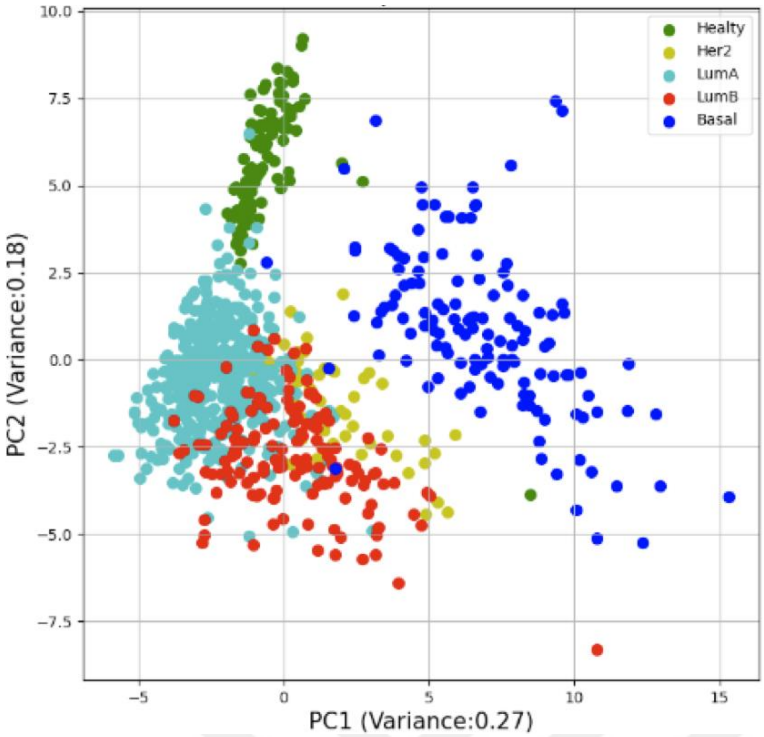


Figure 3.6. PCA Analysis of All Subtypes

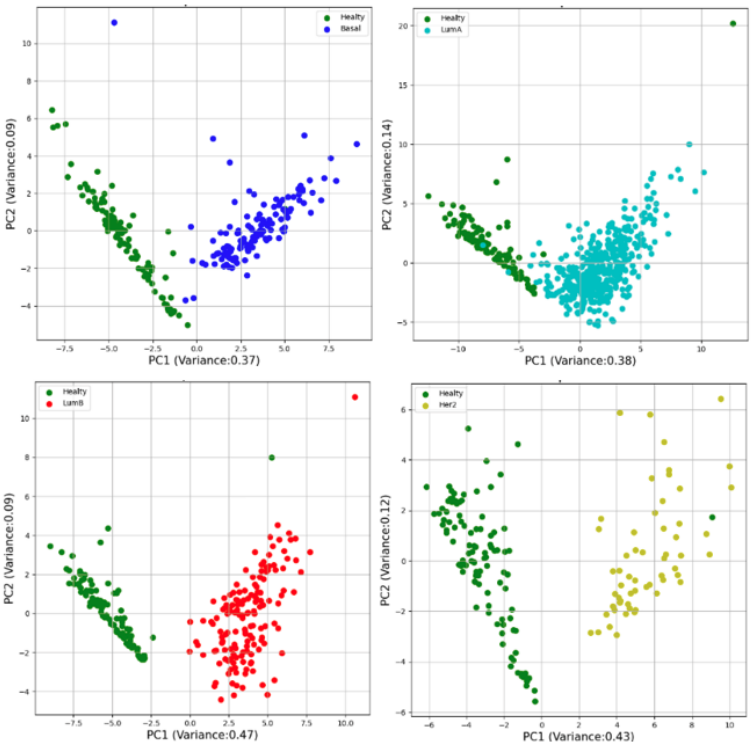


Figure 3.7. PCA Analysis Normal (Healthy) vs Subtype

3.5. Biological Interpretation

The genes encoding the mRNAs present in the RFE_50 model (Figure 3.8) were further investigated through functional enrichment and overall survival analysis. Gene set (Table 3.9) over-representation analysis revealed that these genes encode signaling and regulatory proteins rather than metabolic enzymes.

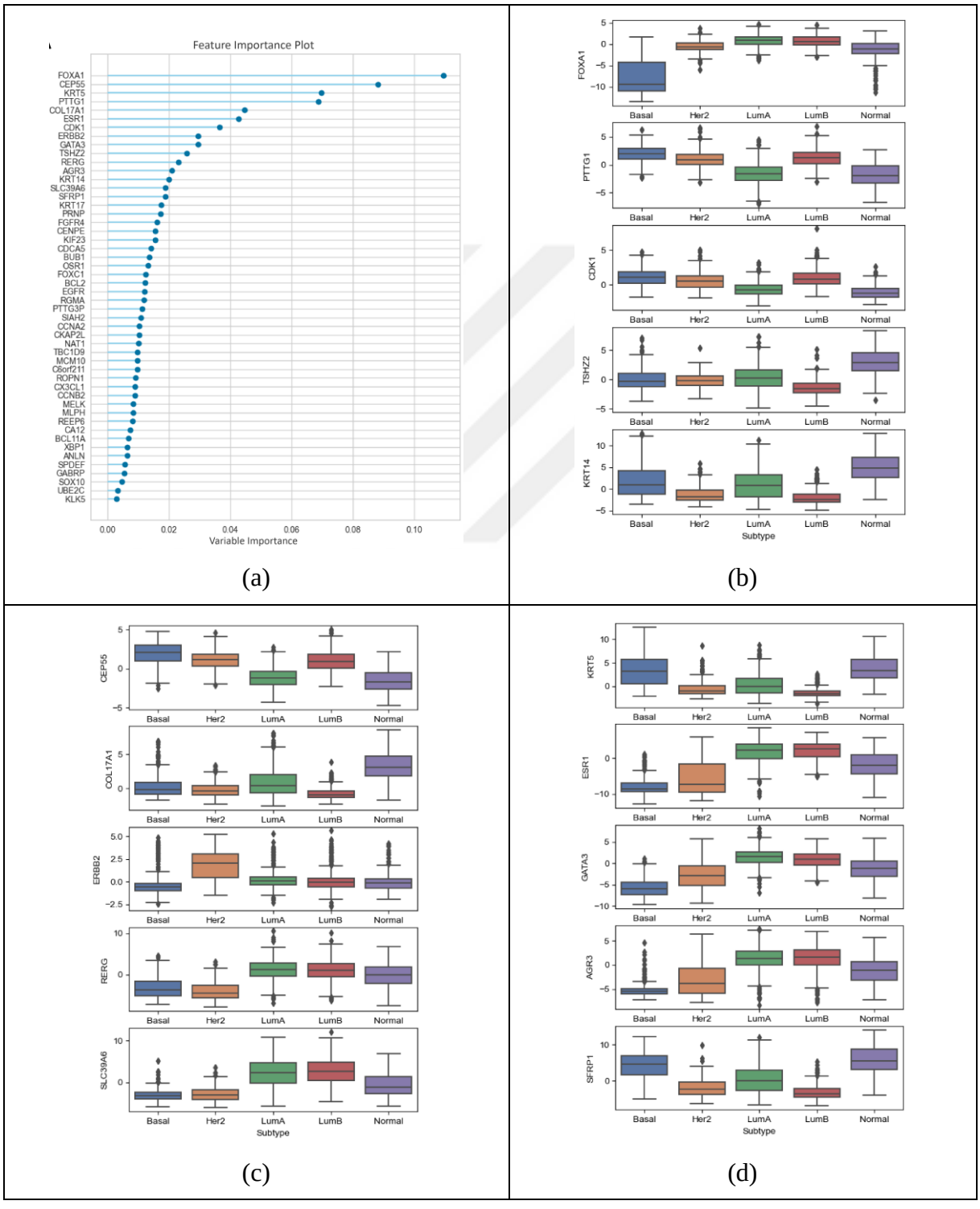


Figure 3.8. RFE_50 Gene Set Over-representation Analysis

Table 3.9. RFE_50 Biomarker Set Selected

RFE_50 Biomarker Set				
NAT1	BCL2	BUB1	CA12	CCNA2
CDK1	CENPE	COL17A1	EGFR	ERBB2
ESR1	FGFR4	FOXC1	GABRP	GATA3
FOXA1	KRT5	KRT14	KRT17	PRNP
CX3CL1	SFRP1	SIAH2	SOX10	XBP1
CCNB2	PTTG1	KIF23	MELK	UBE2C
TBC1D9	SLC39A6	SPDEF	KLK5	PTTG3P
BCL11A	ANLN	ROPN1	CEP55	MCM10
RGMA	MLPH	C6orf211	RERG	REEP6
CDCA5	TSHZ2	OSR1	CKAP2L	AGR3

In addition, pathways and processes that enable the acquisition or maintenance of cancer hallmarks were highlighted. These include cancer pathways such as pathways in cancer, miRNAs in cancer, and breast cancer, signaling pathways such as p53 and PI3K-Akt, and biological processes such as cell cycle, cell death, cell proliferation, cell motility, cellular homeostasis, and focal adhesion (Table 3.10).

Table 3.10 RFE_50 Dataset Biological Interpretation

Signaling pathway	**adj-p-value**	**number of members**
Estrogen signaling pathway	0.001027	KRT17; KRT14; BCL2; EGFR; ESR1
p53 signaling pathway	0.011248	CDK1; BCL2; CCNB2
HIF-1 signaling pathway	0.018322	ERBB2; EGFR; BCL2
PI3K-Akt signaling pathway	0.044720	ERBB2; FGFR4; EGFR; BCL2
ErbB signaling pathway	0.045560	ERBB2; EGFR
Calcium signaling pathway	0.050374	ERBB2; FGFR4; EGFR

Process	**adj-p-value**	**number of members**
Cell cycle	0.001027	PTTG1; CCNA2; BUB1; CDK1; CCNB2
Progesterone-mediated oocyte maturation	0.002985	CCNA2; BUB1; CDK1; CCNB2
Oocyte meiosis	0.005917	PTTG1; CCNB2; BUB1; CDK1
Parathyroid hormone synthesis, secretion and action	0.018322	EGFR; GATA3; BCL2
Cellular senescence	0.033969	CCNB2; CCNA2; CDK1
Central carbon metabolism in cancer	0.044720	ERBB2; EGFR
Adherens junction	0.044720	ERBB2; EGFR
Focal adhesion	0.044720	ERBB2; EGFR; BCL2
Gap junction	0.045783	CDK1; EGFR
Disease pathway	**adj-p-value**	**number of members**
MicroRNAs in cancer	0.014694	ERBB2; CDCA5; EGFR; KIF23; BCL2
Prostate cancer	0.018153	ERBB2; EGFR; BCL2
Pathways in cancer	0.018322	ERBB2; CCNA2; ESR1; EGFR; FGFR4; BCL2
Bladder cancer	0.026219	ERBB2; EGFR
Breast cancer	0.032292	ERBB2; EGFR; ESR1
Gastric cancer	0.032292	ERBB2; EGFR; BCL2
Endometrial cancer	0.037577	ERBB2; EGFR
Non-small cell lung cancer	0.044720	ERBB2; EGFR
Pancreatic cancer	0.044720	ERBB2; EGFR
Proteoglycans in cancer	0.044720	ERBB2; EGFR; ESR1
Human immunodeficiency virus 1 infection	0.045560	CDK1; CCNB2; BCL2
Human T-cell leukemia virus 1 infection	0.045560	PTTG1; CCNA2; CCNB2
Colorectal cancer	0.045560	EGFR; BCL2
Staphylococcus aureus infection	0.049866	KRT17; KRT14

To assess the prognostic significance of the model genes as potential systems biomarkers, the patient cohort was divided into low- and high-risk groups based on the expression profiles of genes encoding mRNAs in the RFE_50 model, and multivariate survival analysis and risk assessments were performed. We found that the genes in the RFE_50 model had a high impact on overall survival ($p=5x10^{-7}$ and HR=2.25) of patients with breast cancer (Figure 3.9).

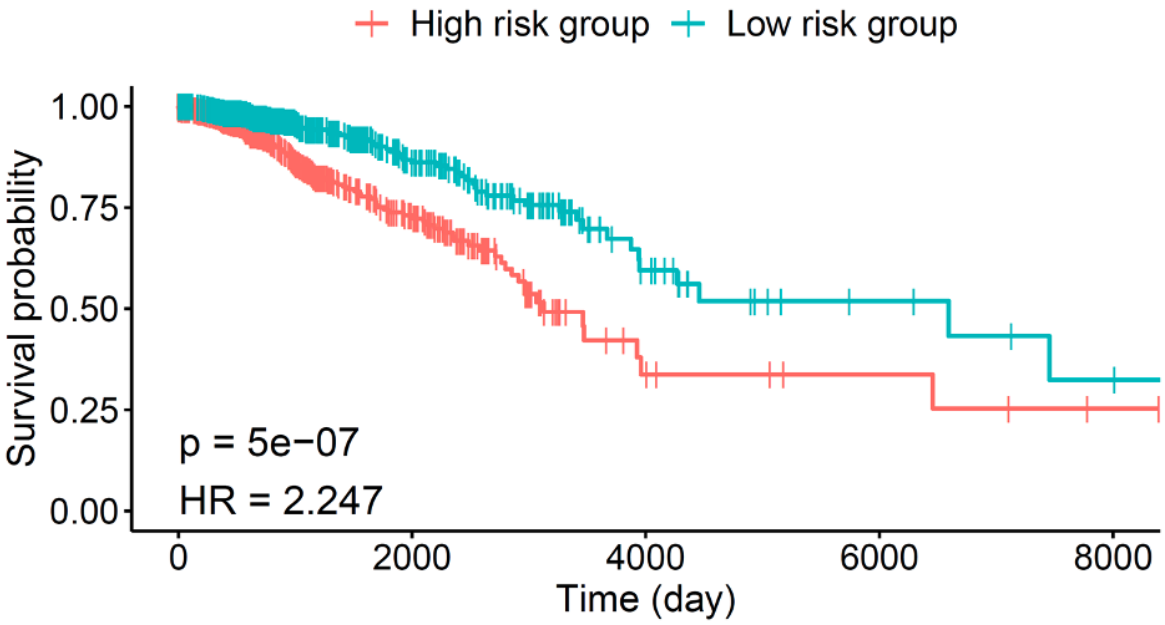


Figure 3.9. RFE_50 Survival Analysis

Precision oncology is tailored to the patient's molecular characteristics to prevent, diagnose, or treat disease, and aims to develop an individualized roadmap for each patient (Yilmaz and Arga, 2022). Precision oncology has made great strides in the last few decades, but developing accurate and computationally intensive algorithms to classify patients to support diagnostic and therapeutic decisions at the point of care is still one of the greatest challenges. Artificial intelligence techniques, particularly ML and deep learning, offer tremendous potential to advance precision oncology (Gulfidan et al., 2021).

Breast cancer is one of the most common malignancies worldwide, with multiple underlying molecular etiologies and subtypes that differ in risk factors, incidence, disease progression, survival, and response to treatment. Thus, there is a clear need for the development of precise strategies to stratify breast cancer into molecular subtypes, not only to assess the prognosis of the disease, but also to enable effective and individualized therapy. Previous studies have shown that ML approaches offer new solutions and avenues for biomedical, bioengineering, and clinical applications (Ozer et al., 2020).

Here, we applied a data mining approach to comprehensive RNA-seq data from the TCGA-BRCA (n=1,215) and METABRIC (n=1,992) projects, hypothesized that changes at the level of gene expression could lead to differentiation of patients with different breast cancer subtypes as well as healthy individuals. By selecting features using alternative feature elimination methods, considering different feature sizes, and following an ensembled learning approach to find the best combination of various ML models, we report here an ensembled model containing 50 mRNAs as features to stratify patients with different breast cancer subtypes as well as healthy individuals. The discriminatory performance of the model reached up to 100% accuracy in pairwise comparisons to distinguish each subtype from the healthy cohort. The features showed high prognostic performance with Hazard ratio of 2.25 ($p=5x10^{-7}$) in breast cancer development and were significantly associated with signaling pathways and biological processes that enable the acquisition or maintenance of hallmarks of cancer.

The main difference and novel aspect of our study is that it is the first study to classify patients with different breast cancer subtypes as well as healthy individuals using an ensembled machine learning approach and utilizing both comprehensive RNA-seq datasets from two large consortia, METABRIC and TCGA-BRCA, for internal and external validation purposes. Recognizing that each modeling algorithm has advantages and disadvantages, we followed the ensembled learning approach, in which six algorithms are used in combination to obtain the best combination of models and followed a two-level blending approach. We reported here an ensembled model including 50 mRNAs as features to stratify patients with different breast cancer subtypes as well as healthy individuals. With up to 100% accuracy in pairwise comparisons for distinguishing each subtype from the healthy cohort, the ensembled learning approach outperformed the other ML algorithms (Ozer et al., 2020).

Our ensembled learning approach provides better performance in stratifying patients into breast cancer subtypes compared with other ML approaches. However, this study also had some limitations. First, although we used the two most comprehensive RNA-seq datasets for breast cancer, the study design was unbalanced with respect to size of patient groups. This has the practical consequence of reducing the statistical power of the models and introducing sampling errors in feature selection. Second, clinical data were limited. Future studies could include more clinical variables in the modeling process. Third, although we evaluated six algorithms in the context of ensembled modeling, there are many algorithms that we did not consider here. Finally, the derived model also needs to be validated using additional independent cohorts at the preclinical and clinical levels.

4. CONCLUSIONS

Considering the need to develop accurate strategies to stratify breast cancer into molecular subtypes, we proposed here an ensemble learning approach to find the best combination of different ML models and presented an ensembled model including 50 mRNAs as features to stratify patients with different breast cancer subtypes as well as healthy individuals.

The performance of the ensemble model and the ML models used are compared, and the ensemble model outperforms the individual ML models. With up to 100% accuracy in pairwise comparisons for distinguishing each subtype from the healthy cohort, the ensemble learning approach outperformed the other ML algorithms.

This study demonstrated that we can consider applying advanced ML methods, such as ensemble learning, to accurately classify cancer patients and provide more powerful tools to improve diagnosis and targeted therapy. Since the amount of training data was small, significant variations were observed when training the models. We recommend further research to improve the performance of our strategy by integrating more clinical and biological data.

REFERENCES

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623-2631).
2. Ashburner M, Ball C, Blake J, *et al.* Gene Ontology: tool for the unification of biology. Nature Genetics, 2000; 25(25–29). https://doi.org/10.1038/75556.
3. Bishop, C. M. (1995). Neural networks for pattern recognition. Oxford university press.
4. Breiman, L. (2001) Random Forests, Machine Learning, 45, 5–32.
5. Chen, T., Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).
6. Chollet, F. (2017). Deep learning with Python. Simon and Schuster.
7. Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., Zmuda, E. (2015). Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, *163*(2), 506-519.
8. Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., Aparicio, S. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature Genetics*, *486*(7403), 346-352.
9. Ferri, F. J., Pudil, P., Hatef, M., & Kittler, J. (1994). Comparative study of techniques for large-scale feature selection. In *Machine Intelligence and Pattern Recognition* (Vol. 16, pp. 403-413). North-Holland.
10. Gatza, M. L., Silva, G. O., Parker, J. S., Fan, C., Perou, C. M. (2014). An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer. *Nature Genetics*, *46*(10), 1051-1059.
11. Gulfidan, G., Beklen, H., Arga, K. Y. (2021). Artificial Intelligence as Accelerator for Genomic Medicine and Planetary Health. *OMICS: A Journal of Integrative Biology*, *25*(12), 745-749.
12. Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K. (2003, November). KNN model-based approach in classification. In *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems",* pp. 986-996. Springer, Berlin, Heidelberg.
13. Guyon, I., Weston, J., Barnhill, S., Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*(1), 389-422.

14. Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, 186(1007), 453-461.

15. Johnson, K. S., Conant, E. F., Soo, M. S. (2021). Molecular subtypes of breast cancer: a review for breast radiologists. *Journal of Breast Imaging*, *3*(1), 12-24.

16. Kamburov, A., Stelzl, U., Lehrach, H., Herwig, R. The ConsensusPathDB interaction database: 2013 update (2013) Nucleic Acids Res, 41, pp. *D793-D800*.

17. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, *45*(D1), D353-D361.

18. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*.

19. Kuhn, M., & Johnson, K. (2019). Feature engineering and selection: A practical approach for predictive models. CRC Press.

20. McLachlan, G. J. (2005). *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons.

21. Montazeri, M., Montazeri, M., Montazeri, M., & Beigzadeh, A. (2016). Machine learning models in breast cancer survival prediction. *Technology and Health Care*, *24*(1), 31-42.

22. Ozer, M. E., Sarica, P. O., Arga, K. Y. (2020). New machine learning applications to accelerate personalized medicine in breast cancer: rise of the support vector machines. *Omics: A Journal of Integrative Biology*, *24*(5), 241-246.

23. Pandala, S.R. Lazy Predict. Available from: https://github.com/shankarpandala/lazypredict. [Last accessed: 05.29.2022].

24. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, E. (2011). SelectKBest, Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825-2830. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html [Last accessed: 05.29.2022].

25. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, E. (2011). Voting Classifier, Scikit-learn: Machine learning in Python. *The Journal of machine Learning research*, *12*, 2825-2830. Available from: https://scikit-

learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html [Last accessed: 05.29.2022].

26. Pfeiffer, R. M., Webb-Vargas, Y., Wheeler, W., Gail, M. H. (2018). Proportion of US trends in breast cancer incidence attributable to long-term changes in risk factor distributions. *Cancer Epidemiology, Biomarkers & Prevention*, *27*(10), 1214-1222.

27. Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers, 10(3), 61-74.

28. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31.

29. Raschka, S., Mirjalili, V. (2019). Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2. Packt Publishing Ltd.

30. Samuel, A. L. (1959). Machine learning. The Technology Review, 62(1), 42-45.

31. Siegel, R.L., Miller, K. D., Fuchs, H. E., Jemal, A. (2022) Cancer statistics. CA Cancer J Clin. 2022;72.

32. Turkki, R., Byckhov, D., Lundin, M., Isola, J., Nordling, S., Kovanen, P. E., Linder, N. (2019). Breast cancer outcome prediction with tumour tissue images and machine learning. *Breast Cancer Research and Treatment*, *177*(1), 41-52.

33. Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Stuart, J. M. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, *45*(10), 1113-1120.

34. Wolpert, D. H. (1992). Stacked generalization. Neural networks, 5(2), 241-259.

35. Wu, J., & Hicks, C. (2021). Breast cancer type classification using machine learning. *Journal of Personalized Medicine*, *11*(2), 61.

36. Yılmaz, B.K., Arga, K.Y. (2022). OMICS: A Journal of Integrative Biology, 358-359.http://doi.org/10.1089/omi.2022.0049.