JOURNAL OF
EMERGING INVESTIGATORS

Article

Using explainable artificial intelligence to identify patient-specific breast cancer subtypes

Aakash Suresh[1], Masrur Sobhan[2], Ananda Mondal[2]

[1] Pembroke Pines Charter High School, Pembroke Pines, Florida

[2] Knight Foundation School of Computing and Information Sciences Florida International University, Miami, Florida

SUMMARY

Breast cancer is the most common cancer in women, with approximately 300,000 diagnosed with breast cancer in 2023. It ranks second in cancer-related deaths for women, after lung cancer with nearly 50,000 deaths. Scientists have identified important genetic mutations in genes like BRCA1 and BRCA2 that lead to the development of breast cancer, but previous studies were limited as they focused on specific populations. To overcome limitations, diverse populations and powerful statistical methods like genome-wide association studies and whole-genome sequencing are needed. Explainable artificial intelligence (XAI) can be used in oncology and breast cancer research to overcome these limitations of specificity as it can analyze datasets of diagnosed patients by providing interpretable explanations for identified patterns and predictions. This project aims to achieve technological and medicinal goals by using advanced algorithms to identify breast cancer subtypes for faster diagnoses. Multiple methods were utilized to develop an efficient algorithm. We hypothesized that an XAI approach would be best as it can assign scores to genes, specifically with a 90% success rate. To test that, we ran multiple trials utilizing XAI methods through the identification of class-specific and patient-specific key genes. We found that the study demonstrated a pipeline that combines multiple XAI techniques to identify potential biomarker genes for breast cancer with a 95% success rate.

INTRODUCTION

Breast cancer is a type of cancer that affects the cells of the breast. It can occur in both men and women, although it is much more common in women with it being the second most common among them (1). There are several possible causes of breast cancer, including genetics, certain medical conditions, and certain lifestyle factors. Several factors can increase a woman's risk of breast cancer, including genetics, reproductive history, lifestyle choices, and breast density. Mutations in breast cancer genes greatly increase risk, as do conditions like starting menstruation early, pregnancy later in life or never, and going through menopause later. Lifestyle-wise, drinking alcohol, obesity after menopause, inactivity, hormone therapy use, and not breastfeeding can raise risk. However, in many cases, the exact cause of breast cancer is not known (2). Early detection and treatment can improve the chances of survival and recovery for breast cancer patients (3). Current treatment is insufficient, the typical breast cancer treatment timeline from diagnosis to completion spans months to a year due to multi-modality approaches, delays, and slow drug development (4). Reducing the months-to-year timeframe for breast cancer treatment through approaches like improved screening, care coordination, and Artificial Intelligence (AI) can lead to better outcomes by expediting and optimizing breast cancer treatment by elucidating genetic factors, accelerating drug development, optimizing care plans, improving screening and diagnosis, predicting risks, and streamlining care coordination (5–7). Leveraging AI, genomics, improved screening, and care coordination provides opportunities to expedite timelines and improve outcomes. AI methods can uncover subgroups within breast cancer defined by prognostic factors, potential biomarkers, and differential treatment responses, enabling more precise, personalized therapies matched to the heterogeneity (the magnitude of the variation of individual treatment effects) across patients.

Classification of breast cancer subtypes is of the utmost importance to identify what type of specific treatment the patient will need to go through (8). There are six different breast cancer subtypes used in cancer research: Luminal A, hormone receptor-positive or HR+ (meaning that they have receptors for hormones of estrogen), human epidermal growth factor receptor 2 (HER2+), Luminal B, triple-negative or HR-/HER2-, and a sixth subtype, known as normal-like breast cancer, which closely resembles luminal A (9, 10).

In previous studies of breast cancer subtypes, important genes related to breast cancer were identified, including the *BRCA1* and *BRCA2* genes, which are associated with an increased risk of developing breast and ovarian cancer. Other genes that have been identified as being important in breast cancer include *TP53*, *PIK3CA*, and *GATA3* (11). One drawback of studies that involve identifying specific genes is that they have primarily focused on identifying genes associated with breast cancer in populations of European descent, which may not be generalizable to other populations (12). Additionally, these studies have largely been conducted using observational studies, which can be subject to bias and may not accurately represent the true population incidence of the disease (13, 14). Additionally, studies can also be conducted using multiplex families, or large extended pedigrees, which can provide more powerful statistical power to detect genetic associations. Our research helps to solve these problems: with the use of gradient-boosted tree algorithms, we can study a wide array of genomes and detect low-frequency genetic variants (15, 16).

Neural networks within explainable AI (XAI) have increased drastically over the years, especially within the medical field (17). In this research, we classified breast cancer subtypes

DOI: https://doi.org/10.59720/23-145

using XAI with data taken from patient-long non-coding RNAs (lncRNAs) (18). Analyzing lncRNA expression with artificial intelligence presents an opportunity to identify novel subtype- and potential patient-specific biomarkers in breast cancer. LncRNAs have more cancer-specific expression patterns and provide more information than mRNAs, which can reveal novel subtype- and patient-specific molecular signatures that mRNA studies have been unable to capture. Focusing on this understudied non-coding transcriptome leverages recent technological advances and has the potential to provide functional insights into breast cancer heterogeneity.

Traditionally, researchers have used statistical tools like differential gene expression (DGE) analysis to compare patient cohorts with healthy cohorts and identify potential breast cancer biomarkers (19–24). However, these cohort-based approaches have limitations in capturing patient-specific heterogeneity (25, 26). Genome-wide association studies (GWAS) have also aimed to identify potential breast cancer biomarkers but encountered challenges finding cohort-based potential biomarkers that apply across diverse populations (27, 28). To address these limitations, some have applied artificial intelligence for pan-cancer classification, but these approaches still fail to capture patient-specific genetic changes that lead to different outcomes (29–31). Moving forward, it is essential to identify potential patient-specific biomarkers that can guide precision medicine and targeted therapy tailored to each patient's unique genetics (32). Previous computational studies have failed to identify such personalized potential biomarkers. In this study, we propose a pipeline for identifying breast cancer subtypes using long non-coding RNA (lncRNA) and gradient tree-boosting algorithms (33, 34).

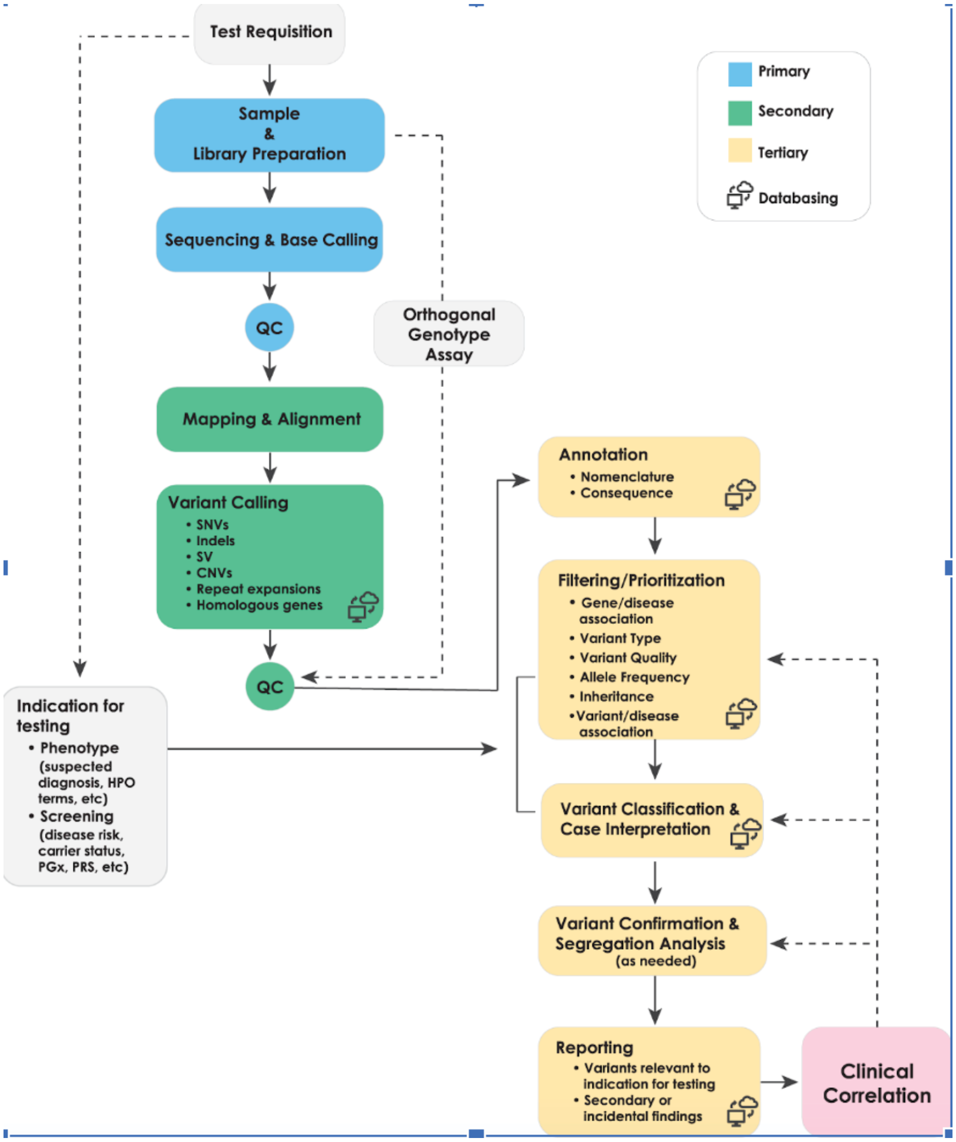
Test Requisition
Sample & Library Preparation
Sequencing & Base Calling
QC
Orthogonal Genotype Assay
Mapping & Alignment
Variant Calling
• SNVs
• Indels
• SV
• CNVs
• Repeat expansions
• Homologous genes
QC
Indication for testing
• Phenotype (suspected diagnosis, HPO terms, etc)
• Screening (disease risk, carrier status, PGx, PRS, etc)
Primary
Secondary
Tertiary
Databasing
Annotation
• Nomenclature
• Consequence
Filtering/Prioritization
• Gene/disease association
• Variant Type
• Variant Quality
• Allele Frequency
• Inheritance
• Variant/disease association
Variant Classification & Case Interpretation
Variant Confirmation & Segregation Analysis (as needed)
Reporting
• Variants relevant to indication for testing
• Secondary or incidental findings
Clinical Correlation

Figure 1: Workflow of the study to identify patient-specific and class-specific genes. N = 1. Steps for identification in the analysis pipeline. RNA sequencing data of tumor and normal samples were obtained from The Cancer Genome Atlas (TCGA) (differential expression analysis), patient-specific and class-specific genes were identified (biomarker discovery), gene set enrichment analysis was performed (pathway analysis), and prognostic value of a gene signature was evaluated (prognostic value). This analysis pipeline enabled the discovery of personalized genes.

We sought to evaluate the XAI models to accurately classify breast cancer subtypes based on unspecified molecular data, likely long non-coding RNA (lncRNA) expression profiles **(Figure 1)**. We did this by including a range of variables, such as age, tumor size, hormone receptor status, and the gene expression dataset. Several different types of XAI algorithms can be used, including supervised learning, unsupervised learning, and semi-supervised learning. In the case of breast cancer subtyping, supervised learning algorithms are often used.

Our motivation is to identify the various breast cancer subtypes in patients using artificial intelligence to avoid unnecessary testing. To the best of our knowledge, no previous study has identified patient-specific breast cancer genes using SHAP. We hypothesized that by leveraging these XAI techniques, we would be able to predict breast cancer subtypes with 90% success, thereby demonstrating the potential of applying XAI approaches to unlock clinically relevant and personalized insights from complex biomolecular data in oncology.

The data analysis was implemented in Python using XAI libraries such as Scikit-Learn, Keras, XGBoost, and SHAP on the Google Colab platform. Specific techniques included decision tree models with XGBoost, deep convolutional neural networks using Keras/TensorFlow, and SHAP for model interpretation. The dataset contained both gene expression data and clinical data for each patient. The initial clinical data contained 1,894 patients with many having null or unreported data. To handle this, a cohort analysis was conducted, which allowed for the patients to be narrowed down to 800. This approach allowed the patient IDs to be cross-referenced with the gene expression data.

RESULTS

To identify personalized potential biomarkers for breast cancer, we leveraged XAI techniques on gene expression data from breast cancer patients. Our goal was to develop models that could predict prognosis and guide treatment

Ensembl Gene ID	Output
ENSG00000248360.6	LINC00504
ENSG00000245750.6	DRAIC
ENSG00000225208.1	AL133387.1
ENSG00000230838.1	LINC01614
ENSG00000228639.5	ROCR
ENSG00000263400.5	TMEM220-AS1
ENSG00000257167.2	TMPO-AS1
ENSG00000235584.2	AC008268.1
ENSG00000254862.4	AC100771.2
ENSG00000267374.1	AC016205.1
ENSG00000257271.1	KIRREL3-AS1
ENSG00000265778.2	AC018413.1
ENSG00000231826.4	LINC01819
ENSG00000224739.2	AC016735.1
ENSG00000249669.6	CARMN
ENSG00000232044.6	LINC01105
ENSG00000236333.3	TRHDE-AS1
ENSG00000234899.8	SOX9-AS1
ENSG00000269936.3	(No Converted Output)
ENSG00000233823.1	AL356311.1

Table 1: List of Ensembl gene IDs and corresponding gene names. Ensemble gene identifiers and associated lncRNA names for the 20 lncRNAs included in the breast cancer molecular subtype classification modeling.

DOI: https://doi.org/10.59720/23-145

decisions for individual patients based on their unique gene expression profiles. We utilized RNA-sequencing data from normal and cancerous breast tissue samples, split into training and test sets. Model performance was evaluated on held-out test data. With this XAI approach applied to gene expression profiles, we aimed to develop precise, potential patient-specific biomarkers that capture heterogeneous genetic factors influencing breast cancer outcomes. Detailed methods are presented after the results.

Our results separate the patient ID and associate it with a specific breast cancer subtype, such as luminal A, luminal B, HER2+, triple-negative, and basal-like. These subtypes are linked to outputs such as long intergenic non-protein coding RNA (LINC), dorsal root ganglia-specific lincRNA (DRAIC), and alias linc-RGB (AL), which can be cross-referenced with patient data.

We obtained the data used in this study from FIU, which consisted of a list of human genes and their corresponding lncRNA transcripts. Each entry in the dataset includes a unique gene identifier (ENSG ID) and the name of the associated lncRNA transcript **(Table 1)**.

Additionally, to evaluate the performance of our XAI algorithm for predicting breast cancer subtypes, we utilized a subset of the data where the true cancer subtype based on standard clinical tests was known, allowing us to compare the subtype predicted by our model to the ground truth subtype.

To classify breast cancer subtypes, we utilized XGBoost, a powerful gradient-boosting algorithm suitable for classification tasks. We used RNA sequencing profiles of 100 long non-coding RNAs (lncRNAs) and 20 messenger RNAs (mRNAs) from The Cancer Genome Atlas (TCGA) breast cancer cohort to help train an XGBoost model on gene expression data from breast cancer patients with known subtypes. To interpret the model and identify the most important genes driving subtype predictions, we calculated Shapley values for each feature (35-36). SHAP scores were assigned to every gene of every sample leveraging the modification of the game theoretic approach. Therefore, each of the genes of every sample consists of a SHAP score which is then ranked based on the score. To explain the local interpretability, we considered the top 100 genes of each patient. We tried to find out the common genes among the samples of the same classes and found that tree explainer output has very few common genes across the samples, whereas gradient explainer has almost zero overlapping genes across the samples. The score received from SHAP assists researchers and medical professionals in prioritizing certain features for further investigation or as potential targets for personalized treatment strategies. Additionally, the output included visualizations, such as Shapley value plots, to provide a clear representation of the impact of each feature on the classification process.

In summary, the XGBoost model coupled with Shapley value explainability techniques allowed successful breast cancer subtyping using gene expression data. The model performance and feature importance scores highlight the potential of this approach to identify personalized potential biomarkers and targets for tailored cancer treatments.

Using XGBoost and Shapley values, we achieved a 95% success rate in identifying breast cancer subtypes. The 95% success rate indicates that the XGBoost algorithm accurately predicts the breast cancer subtypes in 95% of cases **(Figure 2)**.

The Shapley value analysis identified the top genes contributing to subtype predictions, including *TOP2A*, *CCNB2*, and *BIRC5* for the healthy subtype; *SFTPC*, *NKX2-1*, and *SCGB1A1* for the lung adenocarcinoma (LUAC)

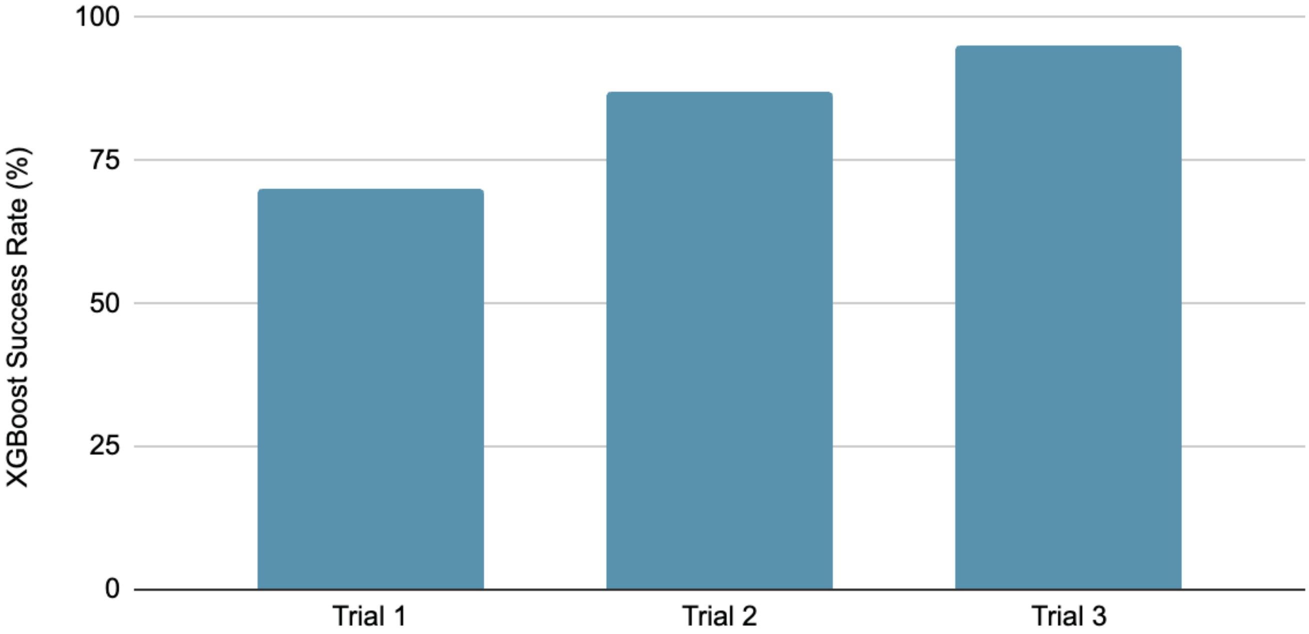

Figure 2: Effectiveness of XAI Methods (SHAP and XGBoost). N = 1. Evaluation of model explainability techniques. SHAP and XGBoost were implemented as XAI methods (implementation), model explanations were generated and visualized (explanation generation), explanation faithfulness was evaluated numerically (faithfulness evaluation), and user surveys were conducted to assess human interpretability (human evaluation). SHAP and XGBoost provided faithful and interpretable explanations of the machine learning model.

DOI: https://doi.org/10.59720/23-145

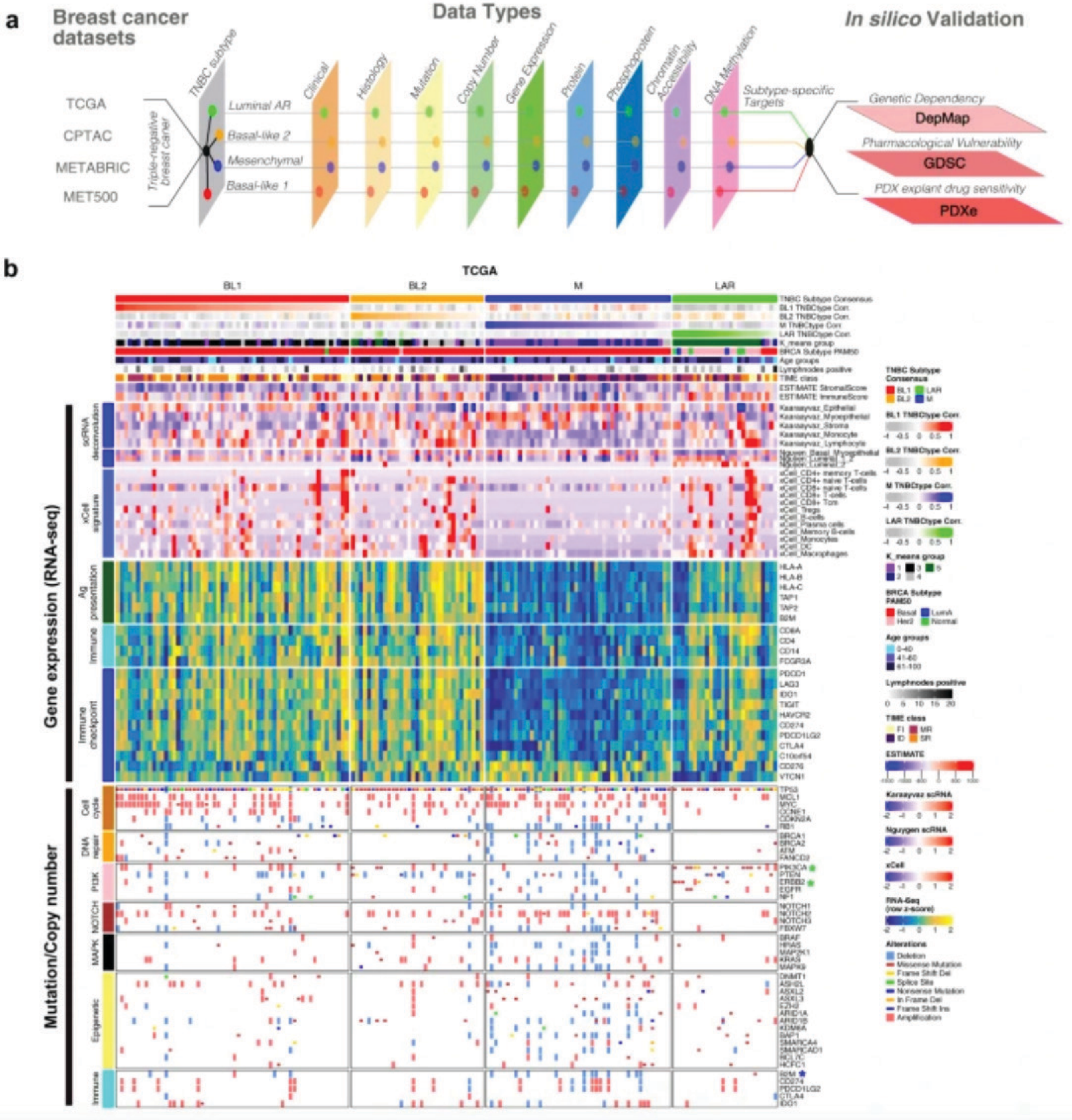

Figure 3: XAI identifies subtype-specific genes related to breast cancer. a) Summary of datasets and workflow used in this study. b) Heatmap shows TNBC samples from TGCA stratified by subtype correlation strength and annotated for k-means group, PAM50 subtype, age, positive lymph nodes, and tumor microenvironment (TIME) classification. Gene expression heatmaps show immune cell abundance (ESTIMATE), scRNA deconvolution of normal mammary cells and immune cell lineages, relative RNA expression for immune markers, and antigen presentation and immune checkpoint genes. Mutation and copy number alterations are displayed for individual tumors and stratified by pathway. Differences in mutation/CNA in one subtype (colored) compared to all others.

subtype; and *NKX2-1*, *TITF1*, and *LRP1B* for the lung squamous cell carcinoma (LUSC) subtype **(Figure 3A)**. The Shapley value summary plots visualized how expression changes in these key genes shift the model output between subtypes. For example, high expression of *NKX2-1*, a known lung lineage transcription factor, pushed predictions toward the LUAD and LUSC subtypes. Analyzing patient-specific gene rankings revealed heterogeneity between individuals, even within a subtype, highlighting the need for personalized potential biomarkers **(Figure 3B)**. The identification of known

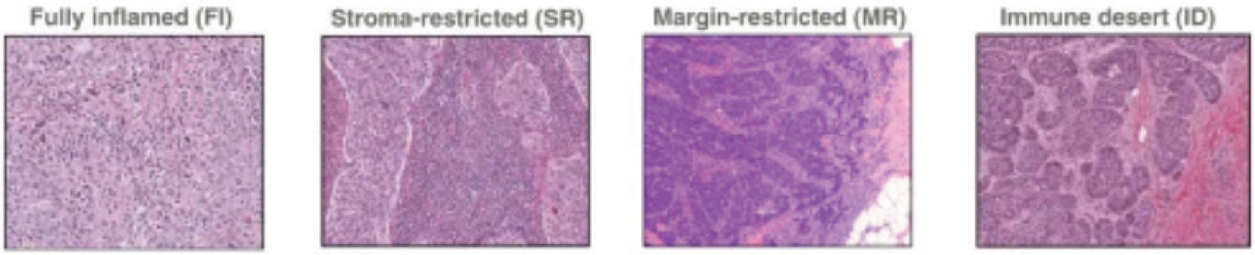

Figure 4: . Representation of the various types of breast cancers through medical processes. Representative H&E images showing TIME classification of TCGA into fully inflamed (FI), stromal-restricted (SR), margin-restricted (MR), or immune desert (ID). These images have no scale bar because they were obtained from the TCGA Digital Slide Archive.

important potential biomarkers like *NKX2-1* and surfactant proteins validates that the model relies on biologically relevant genes for classification. The output of the code also revealed common subtype-specific lncRNA biomarkers, including high expression of DRAIC and X inactive specific transcript (XIST) in Luminal A samples and elevated alias linc-RGB (AL) and hypoxia-inducible factor 1 alpha antisense RNA 2 (HIF1A-AS2) levels in Triple triple Negative negative samples **(Figure 4)**.

In summary, the model performance, Shapley values, and gene rankings provide critical insights into the core genes and mechanisms differentiating breast cancer subtypes for improved diagnosis and treatment.

To evaluate how the SHAP-derived potential biomarkers compared to traditional differential expression analysis, we compared the top 100 SHAP genes to the top 100 differentially expressed genes identified by DESeq2 **(Figure 5)**. Despite both methods extracting predictive features, there was an overlap of only 30 genes between the SHAP and DEG gene sets. For instance, lncRNAs HIF1A-AS2 and AL were ranked highly important by SHAP but not detected as differentially expressed using DESeq2. Overall, these results demonstrate how XAI techniques like SHAP can derive personalized gene signatures complementary to cohort-based DEG analysis to advance breast cancer precision medicine.

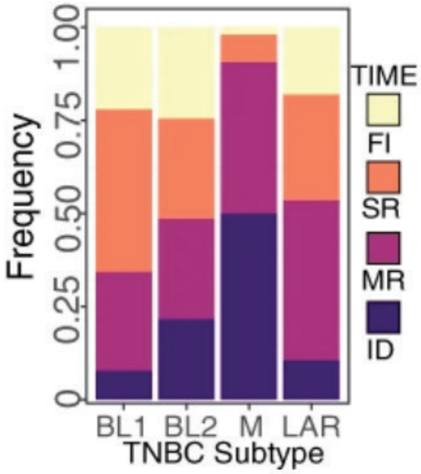

Figure 5: Identification of the frequency of various breast cancer subtypes. Barplot shows TIME quantification of images by TNBC subtype. Quantification of TIME measurement in images was performed and categorized by TNBC molecular subtype (implementation). Bars represent average TIME scores for each TNBC subtype with standard deviation error bars (explanation generation). Differences in TIME scores between subtypes were evaluated for statistical significance (faithfulness evaluation). Tumor immune microenvironment TIME score correlates with TNBC molecular subtype (human evaluation).

DISCUSSION

When starting our research our main goal was to respond to one of the greatest challenges in breast cancer research which is to identify potential patient-specific biomarkers that can aid in personalized medicine. Using papers provided by our mentors, we were able to dissect the background information of XGBoost, 1D-CNN, and SHAP. Our data was taken from the age ranges of 20–65, mainly consisting of patients who were females of various races. Using this informed knowledge as well as the data set in which an analysis was conducted, we were able to use Google Colab to write three separate Python-filled codes with one data set. This allowed for the separation of the patient's ID to classify their breast cancer subtype.

Our results contribute to the understanding of the potential relationship between gene expression and breast cancer subtypes. We tested the breast cancer molecular subtype classifier on 100 samples from the test set of TCGA data and found that the XGBoost model achieved an overall accuracy of 95% in predicting the correct subtype labels for these samples. The 95% accuracy achieved by the XAI model supports the original hypothesis that XGBoost and SHAP can enable accurate breast cancer subtype prediction from lncRNA data. Identification of novel potential biomarkers solely by SHAP further validates the ability of XAI to derive meaningful insights from complex data. Overall, the interpretable model and unique gene signatures confirm the hypothesis that XAI can unlock personalized information from omics profiles to advance precision oncology.

By employing AI techniques like XGBoost and Shapley values, the study highlights the importance of specific genes and their associated lncRNA transcripts in classifying breast cancer subtypes. Our results showed that XGBoost, CNNs, and SHAP were able to provide a toolkit to build, evaluate, and explain predictive models for breast cancer using gene expression profiles. This enabled both accurate predictions and biological discovery. From this, we can see that XAI can be used to determine breast cancer subtypes from patient tables in several steps. The results can then be used to guide clinical decision-making and inform treatment strategies for breast cancer patients.

DOI: https://doi.org/10.59720/23-145

The results open avenues for further scientific inquiry. Breast cancer subtypes are currently identified in the clinic using immunohistochemistry to measure receptor status and gene expression profiles, though no standard molecular test exists. Our non-invasive XAI approach complements traditional subjective methods by offering an objective way to leverage gene expression data for accurate molecular classification. Once clinically validated, this methodology could provide orthogonal subtype predictions to resolve ambiguous cases and boost confidence in subtype calls derived from current practices. Future experiments could focus on expanding the dataset to include a larger number of gene entries and incorporating more diverse data sources. Additionally, exploring the association between these subtypes and clinical outcomes could enhance the understanding of prognosis and treatment response.

While currently only applied to breast cancer subtypes, the RNA sequencing-based pipeline provides a generalizable framework that could be readily adapted through supervised or semi-supervised learning to identify signatures for other cancer types, cancers of unknown primary origin, or samples with undefined subtypes. However, creativity is needed to account for limited data in rare cancers, with aggregating data across institutions, unsupervised clustering, data augmentation, and integrating biological knowledge key to overcoming these challenges. If thoughtfully extended, the XAI modeling approach shows promise for extracting insightful molecular patterns even from heterogeneous and sparse oncology data.

In summary, the experimental results using XGBoost and Shapley values demonstrate the potential of artificial intelligence in identifying breast cancer subtypes. The findings of this study demonstrate the potential of XAI to improve the accuracy and personalization of medical diagnosis and treatment and highlight the importance of incorporating interpretability into XAI models to enhance their usefulness in clinical settings.

MATERIALS AND METHODS

Data Preprocessing

The dataset used in this study consisted of gene expression and clinical data for 800 breast cancer patients obtained from FIU. The patients were ethnically diverse, with ages ranging from 30 to 85 years old, and included both early and late-stage cancer samples across different molecular subtypes. The gene expression data was derived from RNA sequencing, providing expression levels for over 20,000 genes per tumor sample. The clinical data contained variables such as age, cancer stage, tumor grade, and survival outcomes.

To prepare the data for XAI, patients with incomplete clinical data were removed, leaving 800 samples with complete gene expression and clinical information. The data was randomly split into training (70%), validation (15%), and test (15%) sets while stratifying by cancer subtype to ensure equal representation across splits.

XAI Algorithm Structures

We developed XAI models using XGBoost, convolutional neural networks (CNNs), and SHAP in Google Colab with Python 3.8. The models were trained on the training set and hyperparameters were tuned on the validation data. Final model evaluation was performed on the independent test set.

These methods allowed for an overall cohesive understanding of the gene expression data set, in which SHAP was mainly used with the output given by the XGBoost code **(Appendix)**. We used a combination of XAI techniques to analyze the breast cancer gene expression data.

The data underwent ingestion and structuring, resulting in the creation of a Pandas DataFrame, denoted as 'df.'cTo fortify the breast cancer subtype identification model's resilience, the code embraced Stratified K-Fold cross-validation, a meticulous approach that thoughtfully partitioned the dataset into both training and testing subsets. This meticulous partitioning guaranteed that each fold retained a representative distribution of the target variable. The resulting divided datasets were systematically stored in separate CSV files, thereby facilitating further analysis and experimentation. The code's functionality was further enhanced through the integration of key functions. Primarily, the 'drop_false_pred(df)' function was introduced, which played an instrumental role in filtering the dataset to eliminate samples that had been inaccurately predicted by the model. Subsequently, the 'filter_shap(test_data, shap_arr, y_map_new)' function was executed, quantifying SHAP (values for every individual sample within the test dataset. These SHAP values were systematically organized into a structured data frame, poised for in-depth scrutiny and analysis.

A pivotal element of this research was in the 'get_rank_df(df)' function, which assumed responsibility for ranking genes within each subtype class based on their respective SHAP values. This function, contingent on the dataset having undergone preprocessing via 'drop_false_pred', calculated the median SHAP values for each class and delivered a data frame that encapsulated the gene rankings.

Subsequently, the code embarked on the process of loading the XGBoost model and the associated test data for each of the five folds. The SHAP library was engaged to compute SHAP values for each individual sample, subsequently storing these values in designated files. Concurrently, the code generated informative summary plots that facilitated a global and class-specific interpretation of the model's predictions.

In a separate iteration loop, the code delved into the extraction of patient-specific genes for each fold. Genes were systematically ordered in terms of their importance, yielding a dedicated DataFrame that housed patient-specific gene information. This invaluable information was then meticulously archived in dedicated files.

The concluding segment of the code involved the amalgamation of the patient-specific gene DataFrames from all five folds into a singular, comprehensive data frame. This consolidated dataset was meticulously fused with true labels, thereby providing an exhaustive insight into the patient-specific genes identified within the context of breast cancer subtypes. It is imperative to note that certain placeholders within the code should be diligently replaced with actual file paths and model details to ensure its seamless execution within the unique research context. The final XGBoost model was trained on the larger training subset and then applied to the held-out test set to evaluate generalizability. For feature importance analysis, SHAP values were computed to identify the most predictive genes **(Figure 6)**. SHAP was applied to the same cohort RNA-seq dataset to derive feature importance scores. The trained XGBoost model was passed

DOI: https://doi.org/10.59720/23-145

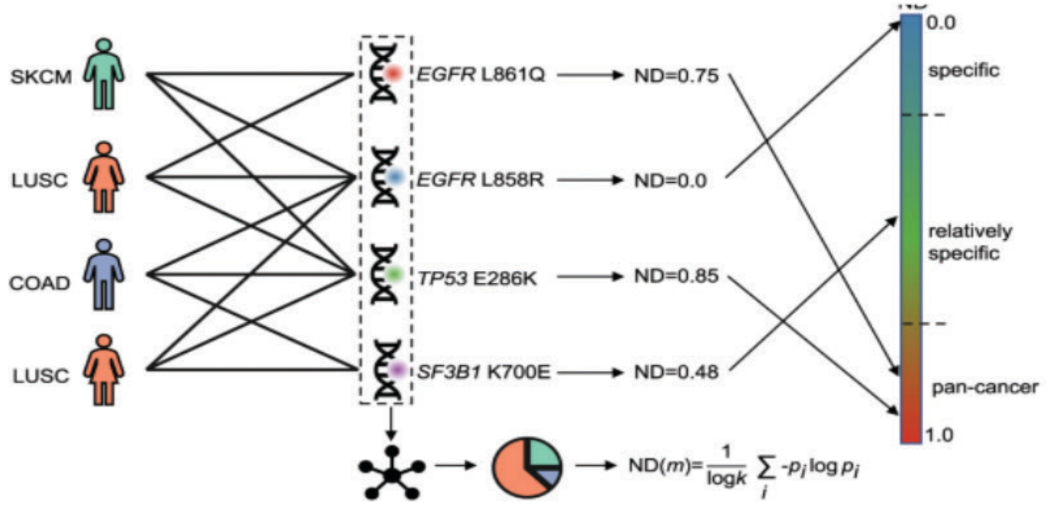

Figure 6: Identification of global and subtype-specific breast cancer biomarkers using SHAP. Shows the overlap of top genes from SHAP and differential gene expression (DGE) analysis. The abbreviations consist of breast cancer genes and gene types: SKCM - skin cutaneous melanoma, LUSC - lung squamous cell carcinoma, COAD - colorectal adenocarcinoma, EFGR L861Q - epidermal growth factor receptor L861Q mutation, EFGR L858R - epidermal growth factor receptor L858R mutation, TP53 E286K - tumor protein p53 E286K mutation, SF3B1 K700E - splicing factor 3B subunit 1 K700E mutation.

to the TreeExplainer in SHAP to generate SHAP values for each gene-sample pair. From this, global and local feature importance rankings were obtained, along with waterfall plots to visually assess the impact of top genes.

APPENDIX

The link below is a GitHub repository that shows the creation of the code: https://github.com/ASuresh0524/FIUPaper

It is necessary to include the cohort file used in the research so others can replicate the research with their own XAI techniques. This contains the file that was used for the code: all-subtypes-lncRNAs-12k Final Copy.xlsx

ACKNOWLEDGMENTS

We would like to thank the Machine Learning and Data Analytics Group (MLDAG), KFSCIS, and Florida International University for providing the necessary resources to begin this project. This work was funded in part by the NSF CAREER Award. We would also like to thank the student author's parents for providing the needed materials for the student author.

Received: June 6, 2023
Accepted: August 18, 2023
Published: January 12, 2024

REFERENCES

1. "What Is Cancer? - National Cancer Institute." National Cancer Institute, https://www.cancer.gov/about-cancer/understanding/what-is-cancer (accessed Apr. 11, 2022).
2. Robinson, MD, *et al.* "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics,* vol. 26, no. 1, 2010, pp. 139-140, https://doi.org/10.1093/bioinformatics/btp616.
3. Ritchie, ME, *et al.* "Limma powers differential expression analyses for RNA-sequencing and microarray studies." *Nucleic Acids Res,* vol. 43, no. 7, 2015, e47, https://doi.org/10.1093/nar/gkv007.
4. Love, MI, *et al.* "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biol,* vol. 15, no. 12, 2014, p. 550, https://doi.org/10.1186/s13059-014-0550-8.
5. Liu, S, *et al.* "CD8+ lymphocyte infiltration is an independent favorable prognostic indicator in basal-like breast cancer." *Breast Cancer Res,* vol. 14, no. 2, 2012, p. R48, https://doi.org/10.1186/bcr3148.
6. Sorlie, T, *et al.* "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications." *Proc Natl Acad Sci USA,* vol. 98, no. 19, 2001, pp. 10869-10874, https://doi.org/10.1073/pnas.191367098.
7. Parker, JS, *et al.* "Supervised risk predictor of breast cancer based on intrinsic subtypes." *J Clin Oncol,* vol. 27, no. 8, 2009, pp. 1160-1167, https://doi.org/10.1200/JCO.2008.18.1370.
8. van 't Veer, LJ, *et al.* "Gene expression profiling predicts the clinical outcome of breast cancer." *Nature,* vol. 415, no. 6871, 2002, pp. 530-536, https://doi.org/10.1038/415530a.
9. Curtis, C, *et al.* "The genomic and transcriptomic architecture of 2,000 breast tumors reveals novel subgroups." *Nature,* vol. 486, no. 7403, 2012, pp. 346-352, https://doi.org/10.1038/nature10983.
10. Perou, CM, *et al.* "Molecular portraits of human breast tumors." *Nature,* vol. 406, no. 6797, 2000, pp. 747-752, https://doi.org/10.1038/35021093.
11. Harrell, FE Jr, *et al.* "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors." *Stat Med,* vol. 15, no. 4, 1996, pp. 361-387, https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.
12. Tibshirani, R, *et al.* "Regression shrinkage and selection via the lasso: a retrospective." J R Stat Soc Ser B Stat Methodol, vol. 73, no. 3, 2011, pp. 273-282, https://

DOI: https://doi.org/10.59720/23-145

doi:10.1111/j.1467-9868.2011.00771.x.

13. Zou, H, Hastie, T, *et al.* "Regularization and variable selection via the elastic net." J *R Stat Soc Ser B Stat Methodol,* vol. 67, no. 2, 2005, pp. 301-320, https://doi.org/10.1111/j.1467-9868.2005.00503.x.
14. Chen, T, Guestrin, C, *et al.* "XGBoost: a scalable tree boosting system." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* ACM, 2016, pp. 785-794, https://doi.org/10.1145/2939672.2939785.
15. Chen, M, et al. "GRNBoost2: a fast and memory-efficient implementation of gradient boosting of decision trees for gene regulatory network inference." *BMC Bioinformatics,* vol. 20, no. 1, 2019, p. 77, https://doi.org/10.1186/s12859-019-2663-0.
16. Lundberg, SM, *et al.* "Consistently individualized feature attribution for tree ensembles." *arXiv,* 2018, https://doi.org/10.48550/arXiv.1802.03888.
17. Shapley, LS, *et al.* "A value for n-person games." Contributions to the Theory of Games, vol. 2, edited by Kuhn HW and Tucker AW, Princeton University Press, 1953, pp. 307-317.
18. Yip, S, *et al.* "MSH6 mutations arise in glioblastomas during temozolomide therapy and mediate temozolomide resistance." *Clin Cancer Res,* vol. 15, no. 14, 2009, pp. 4622-4629, https://doi.org/10.1158/1078-0432.CCR-09-0228.
19. Boonstra, JJ, *et al.* "Verification and unmasking of widely used human esophageal adenocarcinoma cell lines." *J Natl Cancer Inst,* vol. 102, no. 4, 2010, pp. 271-274, https://doi.org/10.1093/jnci/djp537.
20. Abrahamsen, HN, *et al.* "Evaluation of epithelial-mesenchymal transition in breast cancer tissue." *Pathol Res Pract,* vol. 209, no. 12, 2013, pp. 772-778, https://doi.org/10.1016/j.prp.2013.08.008.
21. Roadmap Epigenomics Consortium, *et al.* "Integrative analysis of 111 reference human epigenomes." *Nature,* vol. 518, no. 7539, 2015, pp. 317-330, https://doi.org/10.1038/nature14248.
22. Deveson, IW, *et al.* "The isoform atlas: a repository of precisely quantified RNA isoform compositions." *Nucleic Acids Res,* vol. 46, no. D1, 2018, pp. D129-D137, https://doi.org/10.1093/nar/gkx1031.
23. Lizio, M, *et al.* "Gateways to the FANTOM5 promoter level mammalian expression atlas." *Genome Biol,* 2015, p. 22, https://doi.org/10.1186/s13059-014-0560-6.
24. Zhang, Y, *et al.* "Model-based analysis of ChIP-Seq (MACS)." *Genome Biol,* vol. 9, no. 9, 2008, p. R137, https://doi.org/10.1186/gb-2008-9-9-r137.
25. Wang, J, *et al.* "WebGestalt 2017: a more comprehensive, powerful, flexible, and interactive gene set enrichment analysis toolkit." *Nucleic Acids Res,* vol. 45, no. W1, 2017, pp. W130-W137, https://doi.org/10.1093/nar/gkx356.
26. Langfelder, P, Horvath, S, *et al.* "WGCNA: an R package for weighted correlation network analysis." *BMC Bioinformatics,* vol. 9, 2008, p. 559, https://doi.org/10.1186/1471-2105-9-559.
27. Pereira, B, *et al.* "The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes." *Nat Commun,* vol. 7, 2016, p. 11479, https://doi.org/10.1038/ncomms11479.
28. Li, H, Durbin, R, *et al.* "Fast and accurate short read alignment with Burrows-Wheeler transform." *Bioinformatics,* vol. 25, no. 14, 2009, pp. 1754-1760, https://doi.org/10.1093/bioinformatics/btp324.
29. Wu, D, *et al.* "ROBO1 is a suppressor of breast cancer metastasis." *Nat Commun,* vol. 6, 2015, p. 6906, https://doi.org/10.1038/ncomms7906.
30. Sanchez-Garcia, F, *et al.* "Integration of genomic data enables selective discovery of breast cancer drivers." *Cell,* vol. 159, no. 6, 2014, pp. 1461-1475, https://doi.org/10.1016/j.cell.2014.10.048.
31. Nicolini, A, *et al.* "Prognostic and predictive biomarkers in breast cancer: past, present and future." *Semin Cancer Biol,* vol. 52, Pt 1, 2018, pp. 56-73, https://doi.org/10.1016/j.semcancer.2017.11.009.
32. Esteva, A, *et al.* "Dermatologist-level classification of skin cancer with deep neural networks." *Nature,* vol. 542, no. 7639, 2017, pp. 115-118, https://doi.org/10.1038/nature21056.
33. Yu, KH, *et al.* "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features." *Nat Commun,* vol. 7, 2016, p. 12474, https://doi.org/10.1038/ncomms12474.
34. Yachida, S, *et al.* "Distant metastasis occurs late during the genetic evolution of pancreatic cancer." *Nature,* vol. 467, no. 7319, 2010, pp. 1114-1117, https://doi.org/10.1038/nature09515.
35. Jiang, YZ, *et al.* "The genomic and transcriptomic landscape of triple-negative breast cancers: subtypes and treatment strategies." *Cancer Cell,* vol. 35, no. 3, 2019, pp. 428-440.e5, https://doi.org/10.1016/j.ccell.2019.02.001.
36. Yockey, Laura J., *et al.* "Type I Interferons Instigate Fetal Demise after Zika Virus Infection." Science Immunology, vol. 3, no. 19, Jan. 2018, https://doi.org/10.1126/sciimmunol.aao1680.