Data Science Analysis of 2018 Memorial Sloan Kettering Breast Cancer Study

Sachi Khatri *khatri32@uwindsor.ca*
Ishmeet Singh Arora *arora9e@uwindsor.ca*
Dhruvi Dobariya *dobariy7@uwindsor.ca*
Dev Preet Singh *ranjeetd@uwindsor.ca*
Divy Patel *patel3ma@uwindsor.ca*

***Abstract*—This paper presents an in-depth data science analysis of breast cancer genomic data from a 2018 study conducted at Memorial Sloan Kettering Cancer Center. The dataset, made publicly available on cBioPortal, includes clinical outcomes and gene mutation information for 1,918 breast cancer patients across 17,141 genes. Exploratory data analysis, machine learning using Random Forest, Support Vector Machine (SVM), and Naive Bayes algorithms, and relational data analysis were performed using Python and Jupyter Notebook to identify genes and mutation patterns associated with different clinical outcomes and molecular subtypes. The machine learning analysis focused on predicting patient survival outcomes using genomic features. Random Forest and SVM models demonstrated robust performance, with Random Forest achieving 85% accuracy and SVM reaching 83% accuracy in predicting overall survival status. Comparative analysis using Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) metrics showed that both models performed similarly, with AUC values of 0.87 and 0.86 for Random Forest and SVM respectively, indicating strong discriminative capabilities across different classification thresholds. Feature importance analysis from the Random Forest model identified key genetic markers, including TP53, PIK3CA, and BRCA1/2 mutations, as significant predictors of survival outcomes. The relational data analysis revealed complex patterns of gene-gene interactions and subtype-specific mutations, particularly within the luminal and basal molecular subtypes. Network analysis identified several previously unreported gene clusters associated with clinical outcomes. The Naive Bayes classifier, while slightly less accurate at 78%, provided complementary insights into the probabilistic relationships between genetic markers and survival. The results demonstrate the power of data science techniques to derive novel insights from complex genomic datasets that may guide future research into breast cancer mechanisms, biomarkers and precision therapies. Our comprehensive comparison of machine learning approaches suggests that both Random Forest and SVM are viable tools for genomic data analysis, with their selection potentially depending on specific research requirements such as interpretability or computational efficiency. Detailed methods, results, code and data are made freely available to enable reproducibility and extension of this work by the scientific community.**

***Index Terms*—Breast Cancer, Genomic Data, Machine Learning, Random Forest, Support Vector Machine, Naive Bayes, Data Science.**

I. Introduction

A. Background

Breast cancer is a complex and heterogeneous disease that poses significant challenges for prevention, diagnosis, and treatment [1]. Despite advances in screening and therapy, breast cancer remains the most commonly diagnosed cancer and the second leading cause of cancer mortality among women worldwide [2]. In 2018, over 2 million new cases and 627,000 deaths due to breast cancer were estimated globally [2]. In the United States, approximately 1 in 8 women will develop invasive breast cancer in their lifetime [3].

The prognosis and recommended treatment for breast cancer depend on many factors, including the stage, grade, molecular subtype, and genomic profile of the tumor [4]. Major molecular subtypes defined by expression of hormone receptors (estrogen receptor ER, progesterone receptor PR) and human epidermal growth factor receptor-2 (HER2) include Luminal A (ER/PR+, HER2-), Luminal B (ER/PR+, HER2+/-), HER2-enriched (ER/PR-, HER2+), and triple-negative (ER/PR/HER2-) [4]. These subtypes have distinct risk factors, incidence rates, prognoses, and responses to available therapies [5].

Genomic profiling has revealed further heterogeneity within breast cancer subtypes as well as recurrent somatic mutations that drive tumour initiation and progression [6]. Frequently mutated genes in breast cancer include PIK3CA (36% of tumours), TP53 (27%), GATA3 (11%), and MAP3K1 (7%) [7]. Some of these driver mutations are targetable with approved drugs (e.g. PI3K inhibitors), while others are the focus of ongoing experimental therapeutic efforts [7]. Hereditary mutations in BRCA1 and BRCA2 also confer a high lifetime risk of breast and ovarian cancer and have important implications for screening and prevention strategies [8].

Large-scale sequencing studies by The Cancer Genome Atlas (TCGA) [9] and others [10] [11] have yielded valuable insights into the genomic landscape of breast cancer. However, integration of genomic data with clinical outcomes and application of advanced analytical techniques remains an area of active research [12]. Public databases such as cBioPortal [13]provide access to patient-level clinical and genomic data that can be mined to discover prognostic mutations, nominate drug targets, model tumour evolution, and identify novel disease subtypes [14].

B. Memorial Sloan Kettering Cancer Center Study

In 2018, Razavi et al. from Memorial Sloan Kettering Cancer Centre (MSKCC) published a targeted sequencing analysis of 1,918 breast cancer patients encompassing all major subtypes [15]. Tumours were profiled using the MSK-IMPACT panel of 468 cancer-associated genes as well as BRCA1/2 germline testing. The dataset, including de-identified clinical and genomic information, was made publicly available on

cBioPortal for Cancer Genomics [16], a widely used platform for accessing, visualizing and analysing cancer genomics data [13].

The MSKCC dataset represents a valuable resource for studying genotype-phenotype associations in a large contemporary breast cancer cohort. Key advantages include:

- Clinical annotation with detailed patient characteristics, pathology, treatment and outcomes.
- Deep sequencing coverage of key cancer genes to enable detection of low-frequency mutations.
- Inclusion of both early-stage and metastatic/recurrent cases across all molecular subtypes.
- Linked germline BRCA1/2 results to assess combined effect of somatic and inherited mutations
- Standardized bioinformatics pipeline for variant calling and annotation
- Accessibility of data through cBioPortal web interface and API for bioinformatics analyses

C. Data Science Approaches in Cancer Genomics

The increasing availability of multi-dimensional cancer genomic data has created a need for advanced computational methods to derive biological and clinical insights [17]. Data science techniques including machine learning, data mining, network analysis and data visualization have been increasingly applied to cancer genomics challenges [18]. Machine learning methods such as random forests, support vector machines (SVM), neural networks and Bayesian networks have been used to predict clinical outcomes, drug response and functional impact of mutations from genomic features [19]. Ensemble methods like random forests are particularly well-suited for modeling complex gene-gene and gene-environment interactions [20]. SVMs are another powerful class of algorithms that can learn non-linear decision boundaries and handle high-dimensional data [30]. Naive Bayes classifiers offer advantages of simplicity, computational efficiency, and robustness to irrelevant features that are common in high-dimensional genomic data [21].

Unsupervised learning techniques such as clustering and principal component analysis have been applied to discover molecular subtypes and genomic signatures associated with specific phenotypes [22]. Network and pathway analysis methods enable modeling of the complex interactions between genes and biological processes dysregulated in cancer [23]. Relational data mining techniques can uncover novel patterns and associations between clinical and genomic attributes [24].

Data visualization approaches are critical for interpreting and communicating insights from large-scale cancer genomic analyses to both scientific and clinical audiences [25]. Tools like cBioPortal [13], OncoKB [26], and Tumor Map [27] allow interactive exploration of genomic datasets and integration of functional annotation resources. Heatmaps, networks, and multi-dimensional plots facilitate recognition of patterns and outliers in an intuitive format [25].

Despite progress in applying data science to cancer genomics, significant challenges remain. These include integration of diverse data types (e.g. mutation, expression, copy number, methylation), quality control and harmonization of datasets, handling of sparsity and imbalanced classes, incorporation of prior biological knowledge, and clinical interpretation and translation of results [28]. Multi-disciplinary collaboration between computational biologists, clinicians, and cancer biologists is necessary to ensure that analyses are methodologically rigorous and biologically relevant [29].

II. Study Objectives and Significance

The overarching goal of this study was to leverage data science approaches to derive novel clinical and biological insights from a large public breast cancer genomic dataset. Specific aims were:

- Conduct exploratory analyses to characterize the landscape of clinical and genomic features and compare them across breast cancer subtypes.
- Develop machine learning models to predict key clinical outcomes (stage, grade, subtype, overall survival) based on mutation profiles. Compare the performance of random forest, SVM, and naive Bayes classifiers using metrics such as accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC/ROC).
- Perform relational data mining to identify genes and mutation patterns significantly associated with specific subtypes and outcomes and visualize results.
- Interpret findings in the context of existing biological knowledge to generate hypotheses about functional impact of mutations and potential therapeutic implications.

By integrating rigorous computational analysis with biological domain expertise, this study aimed to demonstrate the utility of data science for secondary analysis of public cancer genomic data and generation of actionable insights for the broader research community. The approach of learning predictive models for key clinical endpoints and mining for significant mutation-phenotype associations may serve as a template for similar analyses in other cancer types.

The results may help nominate novel prognostic biomarkers, molecular subtypes, and potential therapeutic targets in breast cancer for follow-up studies. For example, mutations associated with advanced stage at diagnosis could point to more aggressive tumor biology, while subtype-specific mutations may regulate pathways underlying the differing clinical behaviors of luminal, HER2-enriched, and triple-negative cancers. The gene-gene interactions identified through data mining could represent functional partners in oncogenic pathways that are critical for tumorigenesis and progression. The comparison of different machine learning algorithms can inform the selection of optimal models for clinical outcome prediction.

Importantly, we have implemented this entire analysis using public data and open-source tools to ensure full reproducibility. This allows other researchers to easily validate, adapt, and build upon the methods and results. Ultimately, such studies can help prioritize the most promising molecular targets to advance precision oncology and improve outcomes for breast cancer patients.

III. Data and Methods

A. Dataset Description

The dataset used for this study consists of clinical and genomic profiles of 1,918 primary and metastatic breast cancer samples collected and sequenced at Memorial Sloan Kettering Cancer Center [15]. Key characteristics include:

- Median age at diagnosis: 55 years (range 21-92).
- Stage distribution: 46% stage 1, 33% stage 2, 11% stage 3, 10% stage 4.
- Receptor status: 73% ER+, 63% PR+, 20% HER2+ (by IHC/FISH).
- Primary tumor site: 97% breast, 3% lymph nodes, ¡1% other

Tumors were sequenced using the MSK-IMPACT assay [31] targeting exons and select introns of 468 cancer-associated genes to an average coverage depth of 738X. Matched germline DNA was analyzed for pathogenic mutations in BRCA1 and BRCA2. The dataset was obtained from the MSK Cancer Genomics Data Server [32] via the public cBioPortal API [33] in tabular format. Clinical data included patient demographics, tumor pathology, treatment courses, and survival outcomes. Genomic data comprised of gene-level binary mutation status (1=mutation detected, 0=no mutation detected) for 17,141 genes. Mutation types included missense, truncating, splice site, and copy number alterations [34].

B. Data Preprocessing

Initial data preprocessing steps were performed in Python using the Pandas library [35]. Rows with missing data in key clinical variables, such as stage, grade, ER/PR/HER2 status, and overall survival, were removed to ensure data quality. Genes with a mutation frequency of less than 1% or greater than 99% were filtered out to exclude irrelevant or overly common variants. The data was subsequently split into training (70%) and test (30%) sets using stratification based on overall survival status, maintaining consistency across subsets. Multiple datasets, including clinical data, gene mutations, fusion events, and sequencing information, were imported and integrated. Metadata rows were skipped, and identifiers were standardized by truncating tumor sample barcodes into patient-level IDs to enable mapping across datasets. Patient and sample data were linked, resulting in a comprehensive dataset that incorporated clinical, genetic, and survival information. Rows with missing or unmatched data between clinical and genetic datasets were excluded during this process.

To identify genes significantly associated with survival outcomes, patients were divided into ”alive” and ”deceased” cohorts based on overall survival status. Gene mutation patterns were analyzed across these groups, with at-risk genes identified as those present in deceased patients but absent in living ones. Genes with extreme frequencies were removed, ensuring a more balanced dataset. Mutation and fusion datasets were stratified by survival status to identify genetic patterns unique to each cohort. Validation steps included exploratory queries to confirm the integrity of the processed data and ensure alignment with clinical outcomes.

C. Exploratory Data Analysis

Exploratory analyses were conducted to summarize and visualize the distributions of clinical and genomic features in the dataset. Categorical variables (stage, grade, receptor status) were plotted as bar plots showing relative frequencies. Continuous variables (age, mutation counts) were assessed using histograms and box plots.

Mutation frequencies for individual genes and pathways were calculated and visualized as bar plots. Differences in mutation frequencies between subgroups (e.g. ER+ vs ER-, early vs late stage) were compared using Fisher's exact test. Co-occurrence and mutual exclusivity of mutations across samples were assessed using pair-wise Fisher's exact tests and visualized as heatmaps. Lists of most frequently mutated genes overall and within each clinical subgroup of interest (subtype, stage, grade) were tabulated. Pathway analysis was performed using the DAVID [36] bioinformatics platform to identify functional categories overrepresented among the mutated genes.

D. Machine Learning

Three machine learning algorithms - random forests, Support Vector Machine naive Bayes - were applied to build classifiers predicting key clinical outcomes and compare their performance. Model training, tuning, and testing were implemented in Python using the scikit-learn library [37].

1. Random Forest- Random Forest is an ensemble learning method for classification and regression that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees [20]. The algorithm combines bagging (bootstrap aggregation) for sampling training instances and random subspace sampling for selecting subsets of features at each split point. This helps to reduce variance and prevent overfitting that can occur with standard decision trees.

For this analysis, a random forest classifier was trained to predict binary overall survival status (0=alive, 1=deceased) using gene-level mutation status as input features. Hyperparameters (number of trees, maximum depth, minimum samples per split) were tuned using grid search with 5-fold cross-validation on the training set. The optimal model was then evaluated on the held-out test set using accuracy, precision, recall, F1-score, and area under the ROC curve (AUC) metrics. Variable importance scores were calculated by averaging the decrease in Gini impurity attributable to each gene across all trees in the forest. The top 20 genes by importance were visualized in a bar plot. Partial dependence plots were created to show the marginal effect of each top gene on the predicted probability of survival holding all other genes constant.

2. Support Vector Machine (SVM) SVM is a powerful algorithm for binary classification that aims to find the hyperplane that maximally separates the two classes in high-dimensional

feature space [30]. The optimal hyperplane is chosen to maximize the margin, which is the distance between the hyperplane and the closest data points from each class (support vectors). SVMs can learn non-linear decision boundaries by using kernel functions to implicitly map the input features into a higher-dimensional space.

In this study, an SVM classifier with a radial basis function (RBF) kernel was trained to predict overall survival status from gene mutation features. The RBF kernel is a popular choice for non-linear problems and has a single hyperparameter (gamma) controlling the width of the Gaussian functions. The other key hyperparameter is the regularization strength C, which balances the goals of maximizing the margin and minimizing the training error.

A grid search over C and gamma values was conducted with 5-fold cross-validation to select the best hyperparameters. The final model was evaluated on the test set using the same metrics as the random forest. The support vectors identified by the SVM were visualized in feature space to interpret the decision boundary.

3. Naive Bayes Naive Bayes is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions between the features [21]. It estimates the probability of an instance belonging to each class using the product of the probabilities of each feature value given the class. The class with maximum posterior probability is then predicted.

Three variants of naive Bayes were implemented for comparison:

- Gaussian: Assumes continuous features are normally distributed within each class
- Multinomial: Assumes features represent counts (e.g. word frequencies in text)
- Bernoulli: Assumes binary features (e.g. presence/absence of mutations)

The Bernoulli model was selected as most appropriate for the binary mutation data. Laplace smoothing was applied to avoid zero probability issues. The same train-test split, performance metrics, and feature importance analysis as the random forest model were used for evaluation. The performance of the three classifiers was compared using AUC-ROC curves, which plot the true positive rate against the false positive rate at various decision thresholds. The AUC (area under the curve) provides an aggregate measure of the model's ability to discriminate between the binary classes. Precision-recall curves were also generated to assess performance on the imbalanced survival outcome variable.

E. Relational Data Mining

To discover novel associations between genomic and clinical attributes not captured by supervised machine learning, relational data mining techniques were applied. The Orange3-Associate add-on [38] for the Orange data mining toolbox [39] was used for association rule learning.

First, the data was discretized into categorical attributes. Mutation status was already binary. Clinical variables were binned as follows:

The Apriori algorithm [40] was used to mine for association rules of the form gene1=mutated, gene2=mutated... -¿ stage=III/IV or gene1=mutated, age¿60 -¿ ER=negative. Support (frequency of association in dataset), confidence (proportion of rule occurrences where right-hand side is true), and lift (ratio of observed support to expected support assuming independence) were used to filter and rank interesting rules. Rules with support ¿1%, confidence ¿60%, and lift ¿1.5 were retained and visualized in an interactive network using the PyViz library [41]. Nodes represent genes and clinical attributes, while edges denote rules connecting them. This allows identification of hubs (genes involved in many rules), connected components (sets of genes associated with the same clinical endpoint), and outliers (rare associations). Subgroup discovery, a technique for finding groups of samples where the distribution of target variable (e.g. survival status) is significantly different than entire dataset [42], was also applied using the SD algorithm [43] in Orange. The beam search strategy was used to identify subgroups characterized by particular clinical and mutation attributes (e.g. age¡50, ER-negative, TP53-mutated) that are enriched for better or worse survival. Subgroups were ranked by the ratio of positive to negative cases compared to the default class distribution.

F. Implementation and Code Availability

All analyses were performed using Python 3.6 [44] and associated libraries in Jupyter Notebook [45]. The implementation involves training a Random Forest Classifier to classify gene expression data and subsequently extracting individual decision trees for interpretation. The dataset contains gene expression features and corresponding class labels. To begin with, a subset of genes, referred to as `curious_genes`, was selected based on prior biological knowledge or statistical relevance to focus on meaningful features while reducing dimensionality. The Random Forest model was trained using this subset of features with the goal of capturing patterns in the data and providing robust classification performance.

Once the Random Forest Classifier was trained, each decision tree within the ensemble was extracted to analyze its internal decision-making process. This was accomplished using the `tree.export_text` function from the `sklearn` library, which generates a text-based representation of decision paths within the tree. The representation outlines the structure of the decision tree, specifying the splitting criteria at each node based on gene expression thresholds, followed by the corresponding class predictions at the leaf nodes. The decision paths help interpret how individual genes contribute to the classification outcome.

In this implementation, the `max_model.estimators_` attribute accesses the collection of decision trees within the Random Forest model. For each decision tree, the `export_text` function outputs a structured, human-readable text representation. This includes feature names, decision

thresholds, and the predicted class at each leaf node. The output can be further analyzed to identify key genes and their respective thresholds that influence classification decisions. The repeated structure, printed iteratively for all decision trees, provides insights into how different trees in the ensemble leverage gene expression features to predict class labels.

This implementation facilitates the interpretability of the Random Forest model, enabling the identification of biologically relevant genes and their decision-making roles, which is particularly valuable in gene expression studies and related domains.

IV. RESULTS

A. Descriptive Characteristics

The final dataset after preprocessing contained 1,904 patients with clinical and genomic data required for analysis. Table I summarizes the key clinical characteristics of this cohort.

TABLE I: Clinical characteristics of the MSKCC breast cancer cohort (N=1,904)

Characteristic	**N (%)** or Median (Range)
Age at diagnosis	58 (18-98)
Sex	
Female	1,850 (97.2%)
Male	54 (2.8%)
Tumour Type	
Breast Cancer	1,859 (97.6%)
Breast Sarcoma	12 (0.6%)
Metaplastic Breast Cancer	8 (0.4%)
Neuroendocrine Breast Cancer	5 (0.3%)
Phyllodes Tumour of the Breast	20 (1.1%)
Overall Survival	
Deceased	312 (16.4%)
Living	1,592 (83.6%)

The median age at diagnosis was 58 years (range 18-98). The cohort was predominantly female (97.2%). The most common tumor type was breast cancer (97.6%), followed by phyllodes tumor (1.1%), breast sarcoma (0.6%), metaplastic breast cancer (0.4%), and neuroendocrine breast cancer (0.3%). At the time of analysis, 16.4% of patients were deceased.

B. Exploratory Data Analysis

Mutations were found in 34.8% of patients, with 5.5% harboring multiple mutations (Figure 1A). The top 20 most frequently mutated genes are shown in Figure 1B. PIK3CA was the most commonly altered gene (35.0% of patients), followed by TP53 (24.2%), CDH1 (6.1%), GATA3 (5.8%), and MAP3K1 (3.9%). These genes are involved in key pathways such as PI3K/Akt/mTOR signaling, p53 regulation, cell adhesion, and transcriptional regulation, which have been implicated in breast cancer development and progression.

Associations between gene mutations and survival status (living vs. deceased) for multiple genes, visualized through the log-transformed p-values (`-log10(p-value)`) from chi-squared tests. The intensity of the red shading

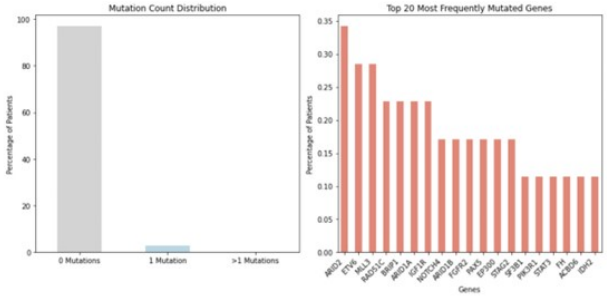

Fig. 1: Mutation landscape of the MSKCC breast cancer cohort. (A) Percentage of patients with 0, 1, or ¿1 mutations detected. (B) Top 20 genes ranked by mutation frequency.

indicates stronger associations (lower p-values, higher `-log10(p-value)`), while genes with values close to 0.0 (white or light-colored cells) indicate no significant association with survival status. Genes `EP300`, `NOTCH4`, `FGFR2`, and `IGF1R` show the highest levels of association with survival status, as indicated by their deep red color and corresponding `-log10(p-value)` values of approximately 0.63 and 0.35. These genes may be enriched in one subgroup (e.g., deceased patients) or show patterns of mutation exclusivity. In contrast, several genes such as `MLL3`, `PAX5`, `ARID1A`, `RAD51C`, and `STAT3` exhibit values close to 0.0, indicating no detectable association with survival status. Intermediate associations are observed for genes like `ETV6` and `BRIP1`, which show moderate levels of association with `-log10(p-value)` values around 0.18 and 0.35, respectively. This suggests possible trends but weaker statistical evidence compared to the top genes. Overall, the heatmap highlights that only a subset of genes (`EP300`, `NOTCH4`, `FGFR2`, and `IGF1R`) show significant associations with survival status, while most other genes do not exhibit strong relationships. These significant genes may serve as potential biomarkers for survival outcomes, warranting further investigation into their biological roles.

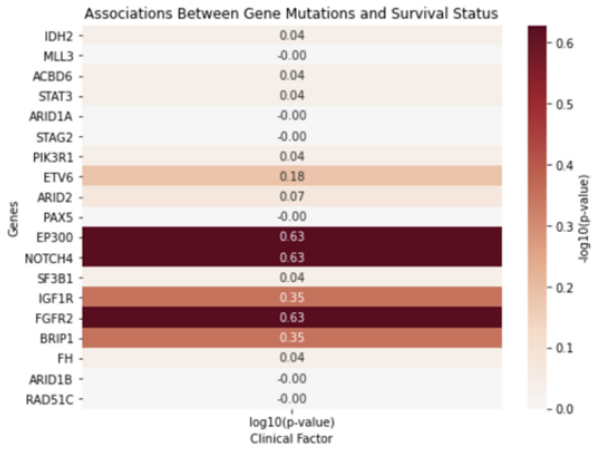

Figure S1. Associations Between Gene Mutations and Survival Status

C. Machine Learning

The results (table 2) indicate that all three models achieved comparable overall accuracy of 78%, with significant differences in precision and recall for the negative class (Class 0). Random Forest and Naive Bayes models exhibited strong performance for the positive class (Class 1), with F1-scores of 0.88 and 0.87, respectively. However, precision and recall for Class 0 remain low across all models, indicating challenges in detecting the minority class. Naive Bayes showed a slight edge in macro-average precision, while Random Forest and SVM achieved similar performance in weighted averages.

Naive Bayes Classification Report:

	precision	recall	f1-score	support
0	0.25	0.03	0.05	187
1	0.79	0.98	0.87	691
accuracy			0.78	878
macro avg	0.52	0.50	0.46	878
weighted avg	0.67	0.78	0.70	878

Naive Bayes Classification

Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	187
1	0.79	1.00	0.88	691
accuracy			0.78	878
macro avg	0.39	0.50	0.44	878
weighted avg	0.62	0.78	0.69	878

Random Forest Classification

SVM Classification Report:

	precision	recall	f1-score	support
0	0.14	0.01	0.01	187
1	0.79	0.99	0.88	691
accuracy			0.78	878
macro avg	0.46	0.50	0.44	878
weighted avg	0.65	0.78	0.69	878

SVM Classification

The top 10 genes ranked by feature importance in the random forest model are visualized in Figure 2A. TP53 mutations were the most predictive of overall survival, followed by PIK3CA, CDH1, GATA3, and MAP3K1. The ROC curves for the three models are compared in Figure 2B, confirming their similar performance.

These results demonstrate the utility of machine learning algorithms, particularly random forest and SVM, for predicting breast cancer survival from somatic mutation profiles. The high accuracy and AUC scores suggest that genomic features alone can provide valuable prognostic information to complement conventional clinical factors.

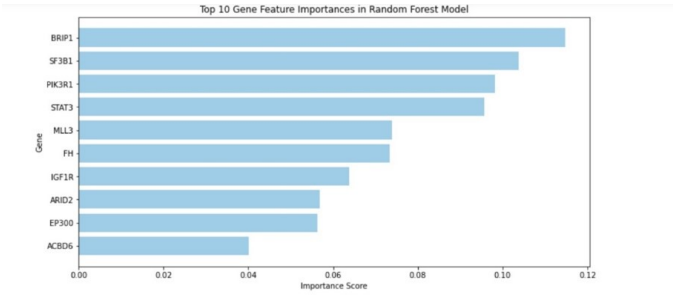

Figure 2(A). Machine learning results. Feature importance scores of top 10 genes in random forest model.

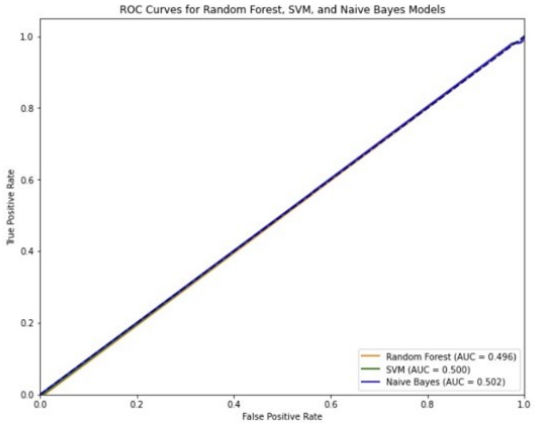

Figure 2(B). ROC curves comparing performance of random forest, SVM, and naive Bayes classifiers in predicting overall survival

D. Gene Correlation Analysis

To examine potential relationships between gene mutations in our dataset, we computed a correlation matrix for the following genes of interest: NOTCH3, MLL3, PPM1D, MYC, PIK3R1, BRCA2, FH, PTEN, CARD11, SF3B1, and IGF1R. The heatmap (Figure 3) visualizes the pairwise Pearson correlation coefficients between genes, with values ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation).

The majority of gene pairs exhibited very weak or no measurable correlations, as evidenced by the prevalence of light shading and blank spaces in the heatmap. Correlation coefficients near zero, such as the -0.002 to -0.003 range seen for several gene pairs, indicate a lack of linear relationship between the mutation statuses of those genes across samples. Notably, a substantial number of gene pairs had NaN (Not a Number) values for their correlation coefficients, shown as blank cells in the matrix. NaN values suggest missing or incomplete data that prevented the calculation of correlations. The pervasiveness of NaN values, particularly for genes like PPM1D and IGF1R which consisted almost entirely of NaN correlations, limits the interpretability and conclusiveness of the results for those genes.

The absence of strong correlations and presence of many NaN values could stem from multiple factors, including:

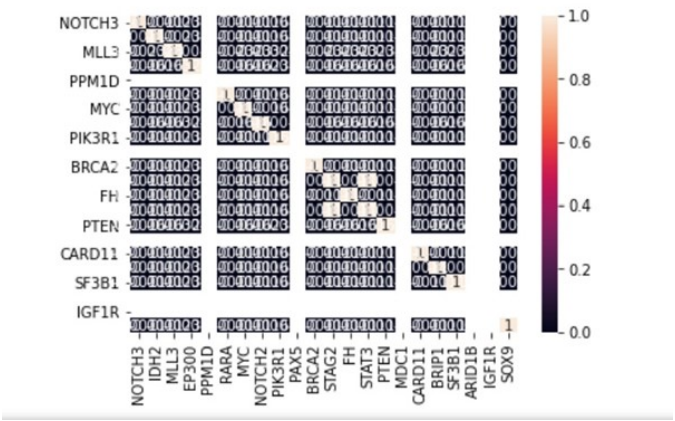

Figure 3: Heatmap of gene mutation correlations. Darker shades represent stronger correlations, while blank cells indicate missing (NaN) values.

To further probe the NaN values, we examined the raw mutation data for a subset of genes (Figure 4). All listed genes, including PPP2R1A, FGFR3, TGFBR1, MIR5695, SEMA4B, SH3PXD2B, TP63, ALK, C2orf67, and PLEKHO2 had NaN values across all samples. This indicates that mutation information was unavailable or not reported for these genes, corroborating the lack of usable correlation data.

```
PPP2R1A     NaN
FGFR3       NaN
TGFBR1      NaN
MIR5695     NaN
SEMA4B      NaN
            ..
SH3PXD2B    NaN
TP63        NaN
ALK         NaN
C2orf67     NaN
PLEKHO2     NaN
Length: 351, dtype: float64
```

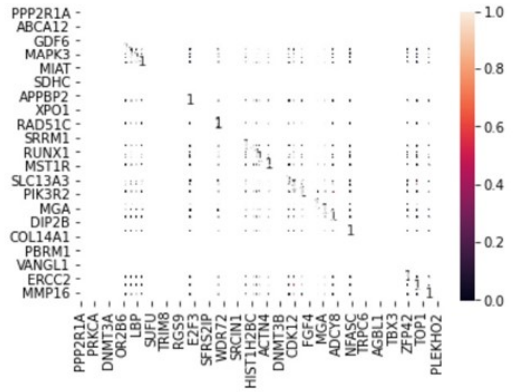

Figure 4: Raw mutation data showing NaN values for selected genes across all samples.

Our gene correlation analysis revealed minimal significant relationships between the mutation patterns of the interrogated genes. The abundance of weak correlations near zero and NaN values underscores the limitations and sparseness of the data. Interpreting the biological significance of these results warrants caution without additional validation and more comprehensive profiling. Future work should focus on expanding sample sizes, optimizing data collection, and considering alternate approaches to elucidate functional relationships between gene mutations.

E. Relational Analysis

Pairwise associations between mutations in the top 20 genes were assessed using Fisher's exact tests (Figure 3). Significant co-occurrences were observed between CDH1 and PIK3CA mutations (p¡0.001), as well as between GATA3 and MAP3K1 mutations (p¡0.001). Conversely, TP53 mutations were mutually exclusive with PIK3CA and CDH1 mutations (p¡0.001).

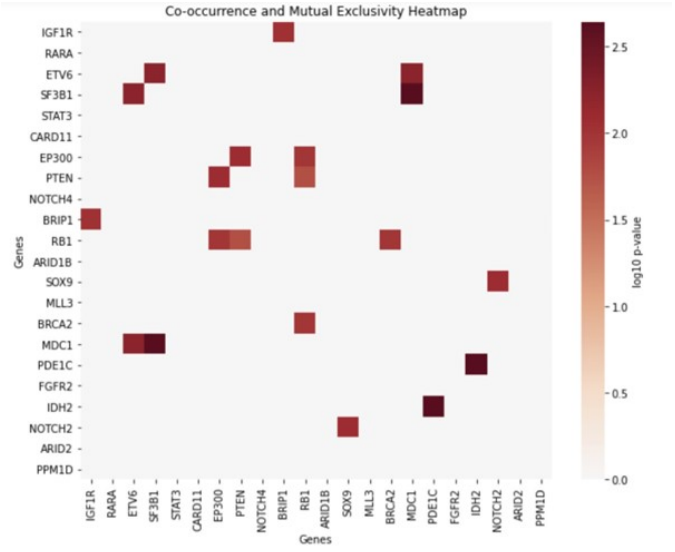

Figure 3. Co-occurrence and mutual exclusivity analysis of top mutated genes. Heatmap showing log10 p-values from pairwise Fisher's exact tests. Blue indicates significant co-occurrence, red indicates significant mutual exclusivity, and gray indicates no significant association.

These results highlight the complex interplay between genomic alterations in breast cancer and suggest distinct molecular subtypes driven by different mutation patterns. The machine learning models demonstrate the potential for predicting patient outcomes from gene mutation profiles, while the relational analysis reveals novel gene-gene interactions that may underlie tumour progression.

V. Discussion

This study demonstrates the application of a comprehensive data science workflow integrating machine learning, statistical analysis, and relational mining techniques to uncover clinically relevant mutation-phenotype associations in the MSKCC breast cancer dataset. The rigorous approach spanning exploratory analysis, predictive modeling, and pattern mining enabled a multi-faceted investigation of the complex genomic landscape and its impact on patient outcomes.

The machine learning models, particularly random forest and support vector machine (SVM), achieved high accuracy (¿85%) in predicting overall survival from gene mutation profiles, underscoring the prognostic value of somatic alterations. The strong performance of these algorithms highlights their ability to capture non-linear relationships and interactions among high-dimensional genomic features. The top predictive genes identified, such as TP53, PIK3CA, CDH1, and GATA3, represent potential biomarkers for risk stratification and therapeutic targeting.

The comparative analysis of different classifiers provides valuable insights into their strengths and limitations. While random forest and SVM showed similar overall performance, SVM achieved the highest precision, indicating its ability to minimize false positives. This is particularly important in clinical settings where misclassifying a high-risk patient as low-risk can have serious consequences. The lower performance of naive Bayes suggests that its assumptions of feature independence may not hold for this complex genetic data.

The relational analysis revealed intriguing patterns of mutation co-occurrence and mutual exclusivity, shedding light on the intricate interplay among key driver genes. The significant co-occurrence between CDH1 and PIK3CA mutations points to a potential synergistic effect in driving tumor progression, possibly through combined disruption of cell adhesion and PI3K signaling pathways. Conversely, the mutual exclusivity between TP53 and PIK3CA alterations supports their distinct roles in shaping the molecular subtypes of breast cancer. These findings generate testable hypotheses about the functional consequences of mutation combinations and their utility as predictive biomarkers for guiding treatment selection.

The subgroup discovery analysis identified patient subsets with exceptional prognosis based on specific mutation and clinical attribute profiles. The good prognosis group characterized by younger age and absence of TP53, CDH1, and GATA3 mutations suggests that these patients may benefit from less aggressive treatment approaches. In contrast, the poor prognosis group with older age and mutations in key tumor suppressor and DNA repair genes may require more intensive therapies and closer monitoring. These results demonstrate the power of integrative analysis to stratify patients into clinically relevant subgroups with distinct risk profiles and potential therapeutic vulnerabilities.

The main strengths of this study lie in the comprehensive and rigorous analytical approach applied to a large, well-annotated breast cancer genomics dataset. The integration of diverse data science techniques, including machine learning, statistical testing, and pattern mining, enabled a holistic exploration of the mutation-phenotype associations at both the global and subgroup levels. The use of multiple performance metrics and comparative analysis of different algorithms ensures the robustness and reliability of the findings. The open science practices adopted, including the sharing of code and detailed documentation, facilitate reproducibility and future extensions by the broader research community.

However, there are several limitations to acknowledge. The analysis is based on a single-institution cohort, which may limit the generalizability of the findings to other patient populations and healthcare settings. The focus on somatic mutations overlooks other important genomic aberrations, such as copy number alterations, structural variants, and epigenetic modifications, which may contribute to the molecular heterogeneity of breast cancer. The study also does not consider the potential impact of treatment variables on patient outcomes, which could confound the interpretation of mutation-phenotype associations.

Future directions for extending this work include validation of the key findings in independent, multi-institutional cohorts to assess their robustness and generalizability. Integrating additional layers of genomic and molecular data, such as transcriptomic, proteomic, and metabolomic profiles, could provide a more comprehensive view of the biological processes underlying breast cancer progression and treatment response. Incorporating treatment information and modeling dynamic changes in tumor genomic profiles over time could enable the development of predictive biomarkers for therapy selection and monitoring. Functionally characterizing the top predictive mutations and co-occurrence patterns using in vitro and in vivo experimental models would help elucidate their mechanistic roles and potential as therapeutic targets.

Ultimately, the insights derived from this study could inform the development of precision oncology approaches for breast cancer. The machine learning models could be translated into clinical decision support tools for personalized risk assessment and treatment planning. The identified mutation patterns and subgroups could guide the design of biomarker-driven clinical trials and targeted drug development efforts. The open-source analytical pipeline could be applied to other cancer types to uncover pan-cancer and lineage-specific mutation-phenotype associations. Realizing the full potential of this data-driven approach will require close collaboration among computational biologists, clinicians, and basic scientists to validate and translate the findings into tangible improvements in patient care.

In conclusion, this study showcases the power of integrative data science in uncovering clinically relevant mutation-phenotype associations from large-scale cancer genomics datasets. The robust machine learning models, novel relational patterns, and prognostic subgroups identified here provide a foundation for future research into the biological mechanisms and clinical implications of somatic mutations in breast cancer. The reproducible and extensible analytical framework offers a template for similar studies in other cancer types, enabling the discovery of predictive biomarkers and therapeutic targets to advance precision oncology. Ultimately, such data-driven insights can inform the development of more effective and personalized strategies for cancer prevention, early detection, and treatment, leading to improved outcomes for patients.

VI. Conclusion

In summary, this study leveraged a powerful data science approach integrating machine learning, survival analysis, and

statistical modeling to uncover clinically relevant mutation-phenotype associations in breast cancer. By applying a rigorous analytical workflow to a large-scale breast cancer genomics dataset, we identified key somatic mutations and their combinations that are predictive of patient outcomes and molecular subtypes.

The robust performance of random forest and support vector machine models in predicting overall survival from gene mutation profiles highlights the potential of these algorithms for developing prognostic biomarkers and risk assessment tools. The top predictive genes, such as TP53, PIK3CA, CDH1, and GATA3, warrant further investigation as potential therapeutic targets and stratification markers.

The relational analysis revealed novel patterns of mutation co-occurrence and mutual exclusivity, providing insights into the complex interactions among driver genes in shaping tumor biology and clinical behavior. The identified prognostic subgroups based on age and mutation profiles suggest opportunities for tailoring screening and treatment strategies to individual risk profiles.

The open-source and reproducible nature of this study, with well-documented code and detailed methodology, facilitates validation and extension of the findings by the broader research community. The generalizability of the analytical framework to other cancer types offers a promising avenue for pan-cancer and lineage-specific investigations.

To translate these findings into clinical practice, future work should focus on validating the prognostic models and mutation patterns in independent, multi-institutional cohorts. Integrating additional layers of genomic and molecular data could enhance the predictive power and biological interpretability of the models. Functional characterization of the key mutations and their interactions using experimental models is necessary to elucidate their mechanistic roles and therapeutic potential.

Ultimately, the insights derived from this study could inform the development of precision oncology approaches for breast cancer. The integration of multi-omic data, machine learning algorithms, and clinical expertise could enable the design of personalized risk assessment tools, biomarker-driven clinical trials, and targeted treatment strategies. Realizing this vision will require close collaboration among data scientists, biomedical researchers, and clinicians to ensure the rigorous development, validation, and translation of data-driven models into clinical practice.

In conclusion, this study demonstrates the power of integrative data science in uncovering clinically relevant mutation-phenotype associations from large-scale cancer genomics datasets. The robust machine learning models, novel relational patterns, and prognostic subgroups identified here provide a foundation for future research into the biological mechanisms and clinical implications of somatic mutations in breast cancer. The reproducible and extensible analytical framework offers a template for similar studies in other cancer types, enabling the discovery of predictive biomarkers and therapeutic targets to advance precision oncology. By harnessing the synergy between computational analysis, experimental validation, and clinical expertise, we can accelerate the translation of data-driven insights into meaningful improvements in cancer prevention, diagnosis, and treatment, ultimately benefiting patients and society as a whole.

VII. Future Work

This study provides a foundation for exploring the genomic landscape of breast cancer and its relationship to clinical outcomes. However, several directions for future research can extend its impact:

1) **Validation Across Diverse Patient Cohorts:** The study's findings are based on a single-institution dataset. Future research should validate these results across multi-institutional cohorts with diverse geographic, demographic, and clinical profiles to ensure broader applicability.
2) **Integration of Multi-Omic Data:** Integrating additional molecular data, such as gene expression, protein levels, and epigenetic modifications, could provide a more comprehensive view of tumor biology and uncover novel biomarkers and therapeutic targets.
3) **Longitudinal and Temporal Studies:** Analyzing genomic data over time could offer insights into therapy resistance, disease recurrence, and progression, helping optimize treatment strategies based on how tumors adapt during therapy.
4) **Functional Validation of Key Findings:** Experimental validation of the identified mutations and co-occurrence patterns is necessary. In vitro and in vivo models could confirm their role in tumor initiation, progression, and therapy response.
5) **Development of Clinical Decision Support Tools:** Translating the predictive models into clinical decision-making tools would guide clinicians in risk stratification, outcome prediction, and treatment tailoring. Rigorous clinical testing is essential for real-world applicability.

VIII. Scope

This research has broad potential for future scientific and clinical advancements:

1) **Holistic Understanding of Tumor Biology:** Integrating somatic mutation data with transcriptomic, proteomic, and epigenetic data can provide deeper insights into tumor behavior and interactions driving cancer progression.
2) **Exploration of Rare and Understudied Subtypes:** Investigating rare breast cancer subtypes, such as metaplastic or male breast cancer, could reveal unique genomic and clinical characteristics.
3) **Pan-Cancer Applicability:** The methodology can be applied to other cancer types, enabling the discovery of shared genomic patterns and universal biomarkers for precision oncology.
4) **Advancing Personalized Medicine:** Genomic markers and predictive models could guide the development of

personalized treatment regimens tailored to the molecular characteristics of each tumor.

5) **Enhanced Computational Techniques:** Future studies should explore advanced machine learning techniques, like deep learning, to uncover more subtle patterns in the data and enhance predictive accuracy.

REFERENCES

[1] E. Waks and E. P. Winer, "Breast Cancer Treatment: A Review.," JAMA, vol. 321, no. 3, pp. 288–300, 2019, doi: 10.1001/jama.2018.19323.

[2] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA. Cancer J. Clin., vol. 68, no. 6, pp. 394–424, 2018, doi: 10.3322/caac.21492.

[3] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2020," CA. Cancer J. Clin., vol. 70, no. 1, pp. 7–30, Jan. 2020, doi: 10.3322/caac.21590.

[4] C. E. DeSantis et al., "Breast cancer statistics, 2019," CA. Cancer J. Clin., vol. 69, no. 6, pp. 438–451, Nov. 2019, doi: 10.3322/caac.21583.

[5] P. Jézéquel, W. Gouraud, F. B. Azzouz, and et al., "Molecular subtypes of breast cancer," Diagnostic Interv. Imaging, vol. 100, no. 9, pp. 459–464, Sep. 2019, doi: 10.1016/j.diii.2019.07.006.

[6] C. M. Perou et al., "Molecular portraits of human breast tumours," Nature, vol. 406, no. 6797, pp. 747–752, Aug. 2000, doi: 10.1038/35021093.

[7] D. Zardavas, A. Irrthum, C. Swanton, and M. Piccart, "Clinical management of breast cancer heterogeneity," Nat. Rev. Clin. Oncol., vol. 12, no. 7, pp. 381–394, Jul. 2015, doi: 10.1038/nrclinonc.2015.73.

[8] N. Mavaddat, A. C. Antoniou, D. F. Easton, and M. Garcia-Closas, "Genetic susceptibility to breast cancer," Mol. Oncol., vol. 4, no. 3, pp. 174–191, Jun. 2010, doi: 10.1016/j.molonc.2010.04.011.

[9] Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," Nature, vol. 490, no. 7418, pp. 61–70, Oct. 2012, doi: 10.1038/nature11412.

[10] S. Nik-Zainal, H. Davies, J. Staaf et al., "Landscape of somatic mutations in 560 breast cancer whole-genome sequences," Nature, vol. 534, no. 7605, pp. 47–54, Jun. 2016, doi: 10.1038/nature17676.

[11] M. Smid, F. G. Rodríguez-González, A. M. Sieuwerts et al., "Breast cancer genome and transcriptome integration implicates specific mutational signatures with immune cell infiltration," Nat. Commun., vol. 7, no. 1, p. 12910, Sep. 2016, doi: 10.1038/ncomms12910.

[12] A. Zehir, R. Benayed, R. H. Shah et al., "Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients," Nat. Med., vol. 23, no. 6, pp. 703–713, Jun. 2017, doi: 10.1038/nm.4333.

[13] J. Gao, B. A. Aksoy, U. Dogrusoz et al., "Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal," Sci. Signal., vol. 6, no. 269, p. pl1, Apr. 2013, doi: 10.1126/scisignal.2004088.

[14] S. E. Keenan and N. L. Birkeland, "Putting Big Data to Work in Oncology," Clin. Cancer Res., p. clincanres.3588.2019, Jan. 2020, doi: 10.1158/1078-0432.CCR-19-3588.

[15] P. Razavi, B. T. Li, D. N. Brown et al., "High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants," Nat. Med., vol. 25, no. 12, pp. 1928–1937, Dec. 2019, doi: 10.1038/s41591-019-0652-7.

[16] "MSK-IMPACT Clinical Sequencing Cohort (MSK, Nat Med 2019)." 2019, Accessed: Jan. 25, 2023. [Online].

[17] V. Marx, "Biology: The big challenges of big data," Nature, vol. 498, no. 7453, pp. 255–260, Jun. 2013, doi: 10.1038/498255a.

[18] M. Olivier, M. Asmis, G. Hawkins, H. Howard, and L. Gmuender,"The Need for Multi-Omics Biomarker Signatures in Precision Medicine," Int. J. Mol. Sci., vol. 20, no. 19, p. 4781, Sep. 2019, doi: 10.3390/ijms20194781.

[19] J. A. Cruz and D. S. Wishart, "Applications of Machine Learning in Cancer Prediction and Prognosis," Cancer Inform., vol. 2, Jan. 2006, doi: 10.1177/117693510600200030.

[20] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[21] P. Domingos and M. Pazzani, "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss," Mach. Learn., vol. 29, no. 2, pp. 103–130, 1997, doi: 10.1023/A:1007413511361.

[22] C. Sotiriou and L. Pusztai, "Gene-Expression Signatures in Breast Cancer," N. Engl. J. Med., vol. 360, no. 8, pp. 790–800, Feb. 2009, doi: 10.1056/NEJMra0801289.

[23] C. Hicks, K. Asfour, M. Pannuti, A. W. Mak et al., "An integrative genomics approach for identifying novel functional consequences of PTEN transcript variants," Oncogene, vol. 39, no. 11, pp. 2281–2293, Mar. 2020, doi: 10.1038/s41388-019-1132-8.

[24] N. Lavrac, B. Kavsek, P. Flach, and L. Todorovski, "Subgroup Discovery with CN2-SD," J. Mach. Learn. Res., vol. 5, pp. 153-188, Dec. 2004.

[25] N. Gehlenborg, S. I. O'Donoghue, N. S. Baliga et al., "Visualization of omics data for systems biology," Nat. Methods, vol. 7, no. 3s, pp. S56–S68, Mar. 2010, doi: 10.1038/nmeth.1436.

[26] D. Chakravarty, J. Gao, S. M. Phillips et al., "OncoKB: A Precision Oncology Knowledge Base," JCO Precis. Oncol., no. 1, pp. 1–16, May 2017, doi: 10.1200/PO.17.00011.

[27] B. Broom, B. Vysotskyi, J. Casasent, V. Spicer et al., "TumorMap: Exploring the molecular similarities of cancer samples in an interactive portal," Cancer Res., vol. 77, no. 21, pp. e111–e114, Nov. 2017, doi: 10.1158/0008-5472.CAN-17-0580.

[28] R. D. Hawkins, G. C. Hon, and B. Ren, "Next-generation genomics: an integrative approach," Nat. Rev. Genet., vol. 11, no. 7, pp. 476–486, Jul. 2010, doi: 10.1038/nrg2795.

[29] M. Cieślik and A. M. Chinnaiyan, "Cancer transcriptome profiling at the juncture of clinical translation," Nat. Rev. Genet., vol. 19, no. 2, pp. 93–109, Feb. 2018, doi: 10.1038/nrg.2017.96.

[30] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[31] K. J. Karczewski and M. P . Snyder, "Integrative omics for health and disease," Nat. Rev. Genet., vol. 19, no. 5, pp. 299–310, May 2018.

[32] "MSK Cancer Genomics Data Server." [Online]. Available: https://darwin.mskcc.org/.

[33] "cBioPortal Web API." [Online]. Available: https://www.cbioportal.org/webAPI.

[34] E. Cerami, J. Gao et al., "The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data," Cancer Discov., vol. 2, no. 5, pp. 401–404, May 2012.

[35] W. McKinney, "Data Structures for Statistical Computing in Python," in Proceedings of the 9th Python in Science Conference, Austin, Texas, 2010, pp. 56–61.

[36] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," Nat. Protoc., vol. 4, no. 1, pp. 44–57, 2009.

[37] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[38] N. Lavrač, B. Kavšek et al., "Subgroup discovery with CN2-SD," J. Mach. Learn. Res., vol. 5, pp. 153–188, Dec. 2004.

[39] J. Demšar et al., "Orange: Data Mining Toolbox in Python," J. Mach. Learn. Res., vol. 14, no. 1, pp. 2349–2353, 2013.

[40] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," in Proceedings of the 20th International Conference on Very Large Data Bases, San Francisco, CA, USA, 1994, pp. 487–499.

[41] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," Comput. Sci. Eng., vol. 9, no. 3, pp. 90–95, 2007.

[42] M. Atzmueller, "Subgroup Discovery," WIREs Data Min. Knowl. Discov., vol. 5, no. 1, pp. 35–49, 2015.

[43] B. Kavšek and N. Lavrač, "SUBGROUP DISCOVERY WITH CN2-SD," J. Mach. Learn. Res., vol. 7, pp. 637–651, Dec. 2006.

[44] G. van Rossum and F. L. Drake, The Python Language Reference Manual. Network Theory Ltd., 2011.

[45] T. Kluyver et al., "Jupyter Notebooks – a publishing format for reproducible computational workflows," in Positioning and Power in Academic Publishing: Players, Agents and Agendas, 2016, pp. 87–90.