**Multi-Model Machine Learning Identifies MAPT, CHEK1, AURKA as Breast Cancer Prognostic Markers**

Jasmine Chan

# Abstract

Breast cancer is the most common cancer among women globally, and the second leading cause of cancer death. While early detection improves survival rates, personalizing treatment based on genomic biomarkers could further increase survival duration. This project applies a multi-model machine learning approach to identify key gene biomarkers correlating with breast cancer survival duration, enabling physicians to personalize treatment. I hypothesized that different models contribute complementary findings: linear regression captures proportional gene-survival relationships, random forest reveals non-linear interactions, and neural network could perform deep analysis on larger datasets but was anticipated to underperform on the small METABRIC dataset (n=1904 patients). Using the METABRIC dataset, I preprocessed gene expression and clinical data by imputing missing values and encoding categorical variables. I trained three models—ElasticNet linear regression, random forest, and PyTorch neural network—using 70/15/15 train/validation/test split to predict overall survival time. Random forest achieved the best test performance ($r^2$=0.147), followed by linear regression ($r^2$=0.138), while the neural network underperformed ($r^2$=0.052) as anticipated. The models identified MAPT, CHEK1, and AURKA as top gene biomarkers strongly associated with survival duration, consistent with published cancer genomics research. The analysis focused on genomic factors, yet survival duration also depends on age, comorbidities, lifestyle, and unrelated causes of death, introducing variance. The consistent under-prediction of survival durations aligns with this limitation and validates the integrity of our genomic-focused approach. These findings demonstrate that complementary models can uncover actionable genomic biomarkers, offering a pathway toward personalized breast cancer treatment and improved prognosis outcomes.

# Introduction

Breast cancer is the most frequently diagnosed cancer among women worldwide, with over 321,910 new cases expected in the United States in 2026 alone (Siegel et al., 2026). It is the second leading cause of cancer death among women. While overall five-year survival rates reach 90% when detected early, outcomes vary dramatically based on cancer subtype, stage, and individual patient characteristics. This variation underscores the urgent need for precision medicine approaches that can predict how long a patient will survive and tailor treatment accordingly.

Existing research in cancer genomics has demonstrated that gene expression profiles can reveal molecular subtypes of breast cancer with distinct clinical outcomes. The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) characterized the genomic

and transcriptomic landscape of over 2,000 breast tumors, identifying novel subgroups with different survival patterns (METABRIC Group et al., 2012; Pereira et al., 2016). However, most machine learning applications in this field focus on predicting binary survival outcomes (alive vs. deceased) rather than continuous survival duration, which is far more clinically valuable for treatment planning.

This project addresses that gap by using three complementary machine learning models to predict survival time and to discover which specific gene biomarkers most strongly influence breast cancer survival duration, enabling physicians to customize treatment plans for individual patients. I hypothesized that different models would contribute multi-dimensional findings. Linear regression identifies genes acting proportionally on survival duration, while random forest captures non-linear and interaction-based relationships for comprehensive coverage. A neural network could perform deep analysis on very large datasets to discover additional differential factors among large populations. However, because the METABRIC dataset contains fewer than 2,000 records, the neural network was anticipated to underperform relative to the other models due to insufficient data volume for deep learning.

# Materials

**Table 1. List of materials used and their purposes.**

| Category | Component / Tool | Specification / Purpose |
|---|---|---|
| **HARDWARE** | **Microsoft Surface Pro 9** | Windows 11 laptop; used for coding, tool access, and results review. |
| | **Google Colab** | Cloud environment; Python T4 GPU for model training and evaluation. |
| **DATASET** | **METABRIC** | 1,904 patients; clinical features and 489 gene expression data (Kaggle). |
| **SOFTWARE** | **Python 3.12.13** | Main programming language for data analysis and ML workflows. |
| | **PyTorch (torch) 2.10.0+cu128** | Deep learning framework for building and training neural networks. |
| | **scikit-learn 1.6.1** | Library for regression, random forest, tuning, and evaluation. |
| | **pandas 2.2.2 / NumPy 2.0.2** | Data cleaning, organization, and numerical operations. |

| | | |
|---|---|---|
| | **SciPy 1.16.3** | Statistical calculations including confidence intervals. |
| | **matplotlib 3.10.0** | Data visualization and graph generation. |
| | **graphviz 0.21** | Decision tree surrogate visualization. |

**Dataset**

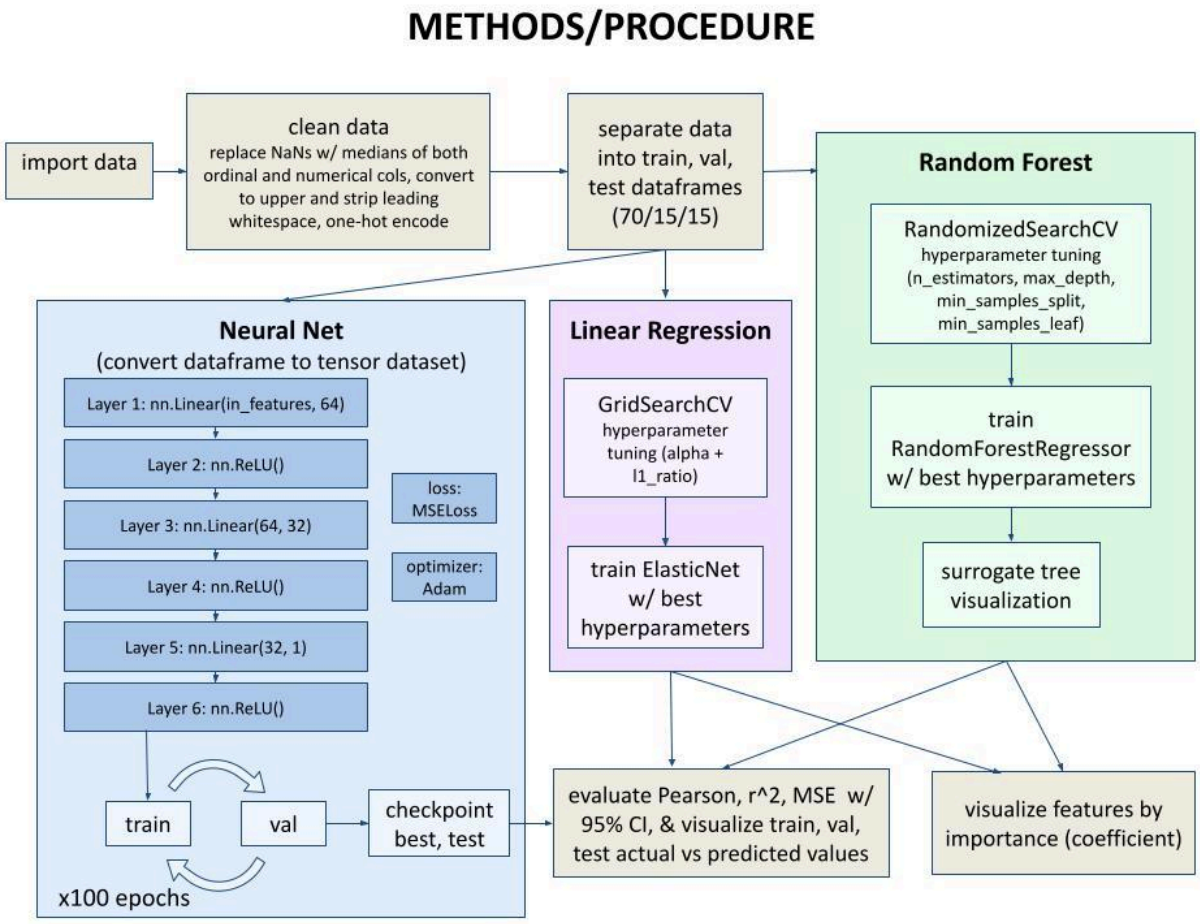
The dataset used in this study was the METABRIC breast cancer dataset, a large public resource for studying breast cancer biology and patient outcomes. As summarized in Table 1, the version used for this project was a Kaggle-distributed dataset (https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric) containing data for approximately 1,900 patients, including clinical variables, gene expression measurements, as well as survival outcomes which were used as prediction targets. The broader METABRIC cohort was originally developed from primary breast tumor samples with long-term clinical follow-up and was designed to support integrated molecular and clinical analysis of breast cancer. In the original METABRIC resource, gene expression profiling was performed using Illumina HT-12 arrays, and the study's initial release included discovery and validation cohorts of 997 and 995 primary breast tumors, respectively. Later METABRIC work also expanded the molecular characterization of the cohort through targeted sequencing of 173 recurrently mutated breast cancer genes, helping establish METABRIC as a widely used benchmark dataset for prognostic modeling and biomarker discovery in breast cancer research.

**Hardware/Software**

The computational resources used in this study are summarized in Table 1. A Microsoft Surface Pro 9 laptop computer (Microsoft Corp., Windows 11) was used for code development, access to online tools, organization of files, and review of model outputs. Data preprocessing, machine learning model training, and evaluation were carried out primarily in the Google Colab cloud notebook environment (Google LLC), which provided access to remote computing resources, including a remote NVIDIA T4 GPU for accelerated model training and testing. Python 3 was used as the main programming language for implementing the analysis workflow. The primary software libraries included PyTorch (torch) for neural network development, pandas for loading and organizing tabular data, NumPy for numerical computation, SciPy for statistical analysis, matplotlib for graph generation and visualization, scikit-learn for machine learning model building and evaluation, and graphviz for decision tree surrogate visualization.

# Methods

**Figure 1. Block diagram describing the procedure used to process the data.**




The overall workflow used in this study is summarized in Figure 1. First, the METABRIC breast cancer dataset was obtained from Kaggle and the raw CSV file was imported into Python via pandas for preprocessing and analysis. Data cleaning was then performed to prepare the dataset for machine learning. A patient ID mapping dictionary was created for row identification, after which the patient id column was removed from the model inputs so that it would not influence prediction. The target variable was defined as time until death in years, calculated by dividing overall survival months by 12. Missing values (NaNs) were assessed and imputed using the median for numerical columns and the mode for categorical columns. To reduce noise and prevent data leakage, outcome-related or otherwise non-informative columns were removed, including overall survival, overall survival months, cohort number, reason for death, and the nottingham prognostic index, a tool that separates patients into clinical groups based on their prognosis (Fong et al., 2015). In addition, categorical values were standardized where needed by cleaning formatting inconsistencies such as data type, capitalization, and extra whitespace. Categorical predictors were one-hot encoded for model use. After preprocessing, the dataset was split into training, validation, and test sets using a 70/15/15 split with a fixed random seed of 42, as shown in Figure 1. The test set was reserved for final model evaluation only, while the training and validation sets were used during model development and tuning.

Three machine learning approaches were trained and compared. The first model was an ElasticNet linear regression model, which combined L1 and L2 regularization. Its hyperparameters, alpha and l1 ratio, were optimized using GridSearchCV. The second model was a Random Forest Regressor, for which the hyperparameters n estimators, max depth, min samples split, and min samples leaf were tuned using RandomizedSearchCV. The third model was a PyTorch neural network consisting of six layers with ReLU activation functions, mean squared error (MSE) loss, and the Adam optimizer. The neural network was trained for 100 epochs, with validation performance monitored during training and checkpointing used to save the best-performing model before final testing. As illustrated in Figure 1, all three models were ultimately evaluated on the same held-out test set.

**Statistical Analysis**

Following training, model performance was compared using Pearson correlation coefficient (r), coefficient of determination ($r^2$), and mean squared error (MSE), each reported with 95% confidence intervals on the held-out test set.

Pearson's r was used to measure the strength and direction of the linear relationship between the predicted survival times and the actual survival times. A value closer to 1 indicates that the predicted and actual values increase together more consistently, while a value closer to 0 indicates little linear relationship. The $r^2$ value was used to measure how much of the variation in survival time was explained by the model. Higher $r^2$ values indicate that the model accounts for more of the observed variability in the outcome, whereas values closer to 0 indicate weaker explanatory power. The MSE was used to quantify prediction error by averaging the squared differences between predicted and actual survival times. Lower MSE values indicate more accurate predictions. To improve the reliability of the results, 95% confidence intervals were also calculated for the test-set performance metrics in order to ensure statistical significance.

In addition to predictive performance, model interpretability was examined by extracting feature importance rankings from each method. For the Random Forest model, a surrogate decision tree visualization was also generated to provide a more interpretable representation of model behavior.

# Results

**Figure 2. Comparison of actual versus estimated actual survival years for models run on the test set.** (a) Linear regression comparison. (b) Random forest comparison. (c) Neural network comparison.

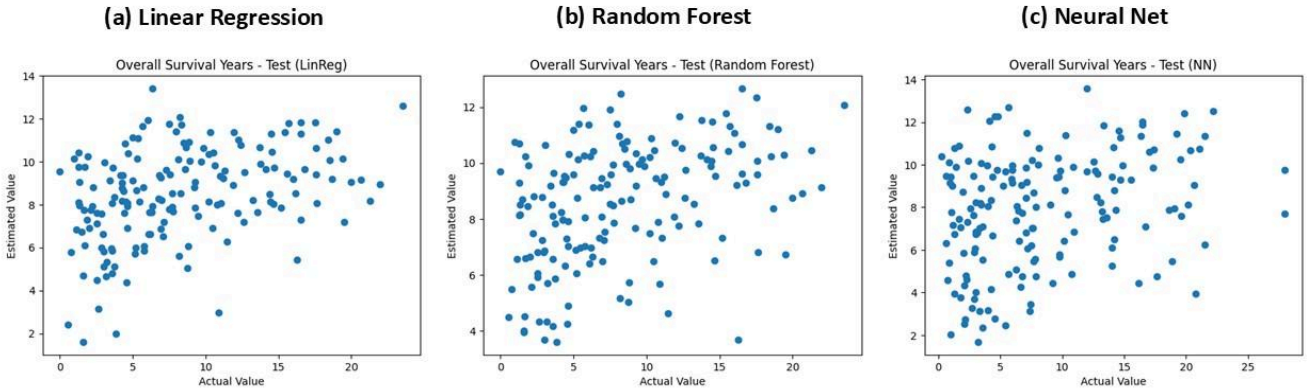



**Table 2. Evaluation metrics for models run on the test set with a 95% confidence interval.** Displays the Pearson r, $r^2$, and mean squared error (MSE) for the (a) linear regression, (b) random forest, and (c) neural net models.

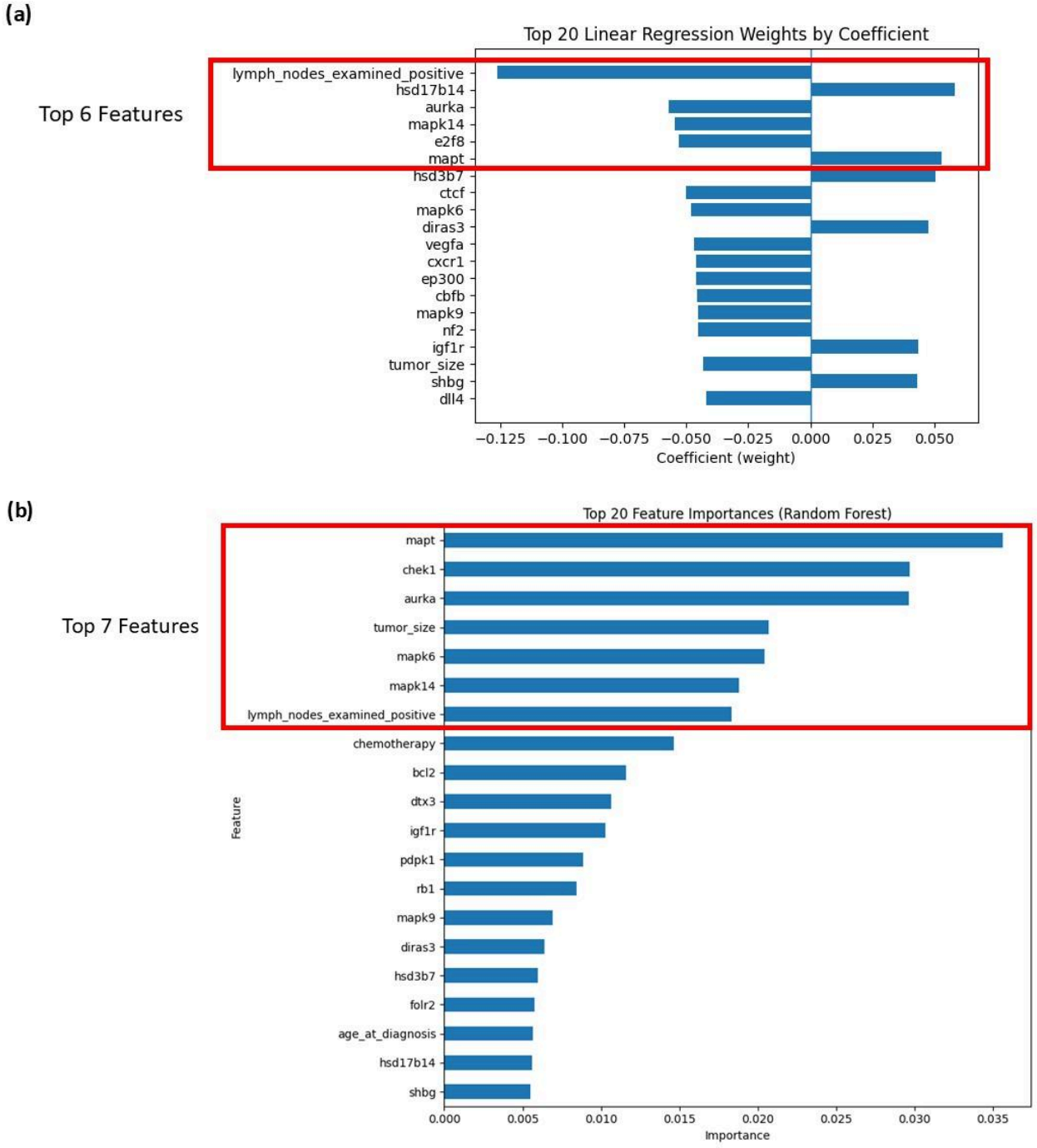
| Metric | (a) Linear Regression | (b) Random Forest | (c) Neural Network |
|---|---|---|---|
| **Test Pearson r** | 0.379 [0.262, 0.492] | 0.392 [0.258, 0.516] | 0.288 [0.148, 0.416] |
| **Test $R^2$** | 0.138 [0.043, 0.227] | 0.147 [0.039, 0.246] | 0.052 [-0.061, 0.161] |
| **Test MSE** | 26.187 [21.222, 31.540] | 25.910 [20.815, 31.362] | 37.414 [29.191, 46.972] |

Each model highlighted different types of relationships between input features and breast cancer survival duration. The linear regression model identified features with approximately linear associations with survival time, with lymph nodes examined positive emerging as the strongest linear predictor. In contrast, the Random Forest model captured non-linear relationships between gene-expression features and survival duration, identifying MAPT (Microtubule-Associated Protein Tau), CHEK1 (Checkpoint Kinase 1), and AURKA (Aurora Kinase A) among the most important predictors. The neural network model was also trained and evaluated, but it did not perform as well as the linear regression and Random Forest models based on the evaluation metrics used in this study.

Among the gene-expression features analyzed, MAPT, CHEK1, and AURKA emerged as the most notable biomarkers associated with breast cancer survival duration. MAPT ranked first in Random Forest feature importance, indicating a strong contribution to the non-linear model.

**Figure 3. Top feature importances.** (a) Top 20 linear regression feature weights, ranked by magnitude regardless of positive or negative direction. Bar direction displays the positive and negative direction of each respective coefficient, with positive coefficients indicating high correlation with survival time and negative coefficients indicating low survival time. (b) Top 20

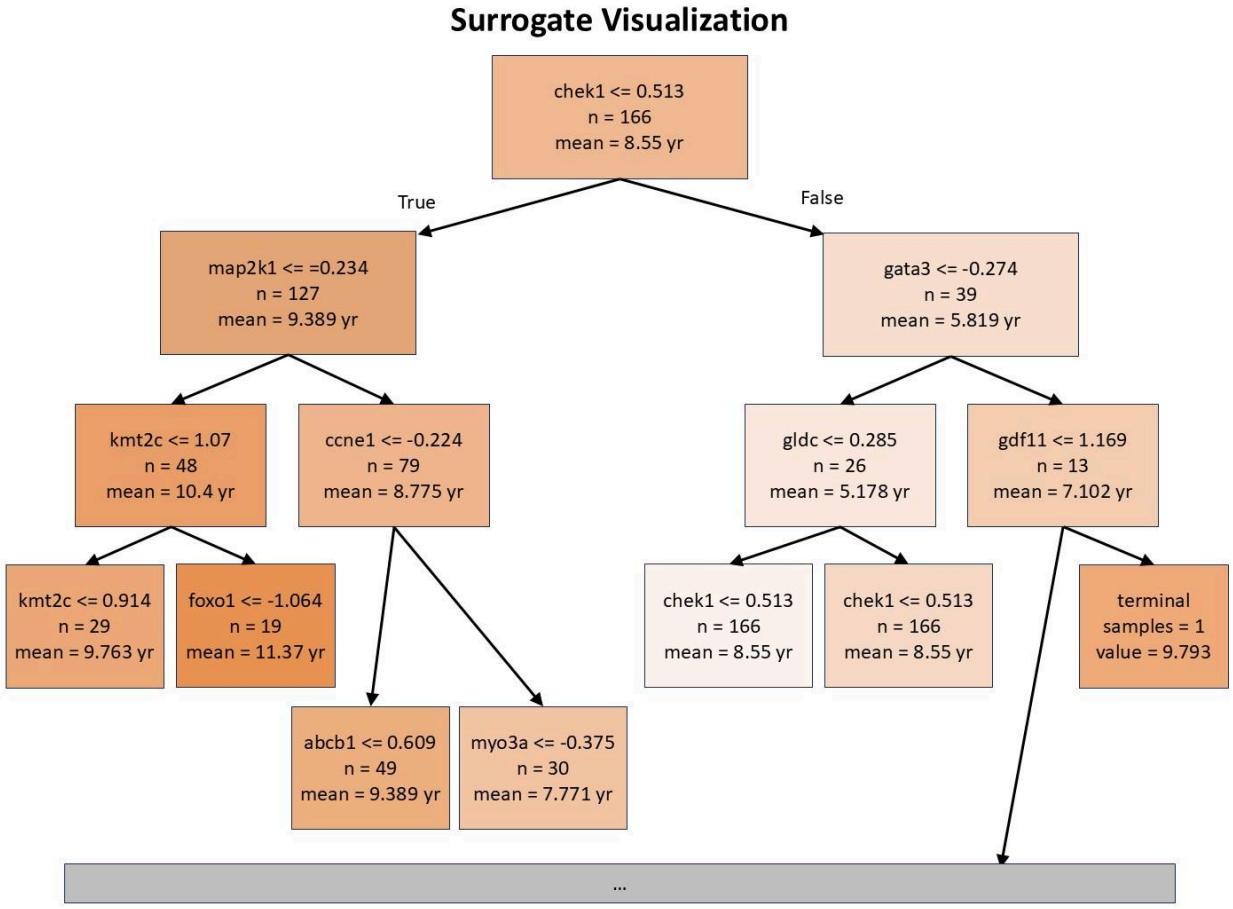
random forest feature weights, ranked by magnitude. Unlike the linear regression features, random forest features display only importance, with higher values indicating more contribution to prediction and lower values indicating less, with no positive or negative correlation known.




CHEK1 appeared as the root node in the surrogate decision tree, suggesting that it served as the primary splitting feature in the interpretable approximation of the Random Forest model. AURKA ranked within the top three features across both the linear regression and Random Forest models, demonstrating consistency across methods. Although lymph nodes

examined to be positive showed the strongest linear association with survival, it is a conventional clinical prognostic variable rather than a gene-expression biomarker, which provides indication that correlations are accurate. In contrast, MAPT, CHEK1, and AURKA were the most prominent genomic features identified in this analysis.

**Figure 4. Surrogate tree visualization.** Displays decision splits made by the model at the first three layers; cut short due to size and visualization purposes.




To improve interpretability of the Random Forest model, a surrogate decision tree was trained to approximate the ensemble's predictions. Because a Random Forest combines many individual decision trees, its internal prediction process can be difficult to interpret directly. The surrogate tree distills the forest's complex decision-making process into a human-readable form, revealing which genes the model considers most important and how it uses them to stratify patients by predicted survival duration. The root node split on CHEK1 at a threshold of 0.513, identifying it as the most important feature for the initial division of patients. Patients with lower CHEK1 expression (CHEK1 ≤ 0.513) had a mean predicted survival of 9.389 years (n = 127), whereas patients with higher expression (CHEK1 > 0.513) had a mean predicted survival of 5.819 years (n = 39). Subsequent splits involved MAP2K1, GATA3, KMT2C, and CCNE1, indicating that the model used a multi-gene decision pathway to further distinguish survival patterns across patient subgroups.

# Discussion

The project was able to utilize a multi-model approach and successfully identified three clinically validated gene biomarkers—MAPT, CHEK1, and AURKA—that correlate with breast cancer survival duration. The convergence of findings across different model types strengthens confidence in these biomarkers: MAPT and AURKA appeared in the top features of both linear regression and random forest, while CHEK1 emerged as the primary decision factor in the surrogate tree visualization. These findings resonate with existing cancer genomics literature. MAPT (Tau protein) expression has been positively correlated with breast cancer survival (Bonneau et al., 2015; Callari et al., 2023). CHEK1 overexpression has been established as an independent risk biomarker associated with aggressive breast cancer phenotype (Wu et al., 2019). AURKA has been shown to outperform Ki67 as a prognostic marker in ER-positive breast cancer (Ali et al., 2012). The alignment between our computational findings and published clinical research validates the multi-model approach.

The analysis focused primarily on genomic and gene-expression factors. However, actual survival duration depends on many additional variables including patient age, comorbidities, lifestyle factors (diet, exercise, emotional health), treatment regimens, and even unrelated causes of death (accidents, other diseases). These non-genomic factors introduce substantial variance into survival prediction. Notably, the models generally predicted survival durations lower than actual values. This under-prediction is consistent with the absence of protective factors in the model—while genomic markers capture risk, real-world outcomes are extended by factors such as effective treatment, healthcare access, and lifestyle variables that were not included. This observation highlights opportunities for future work incorporating broader clinical and demographic data.

The neural network underperformed ($r^2$=0.052) compared to linear regression and random forest, which was anticipated given the METABRIC dataset contains fewer than 2,000 records. Neural networks require substantially larger datasets to learn complex patterns without overfitting. The training/validation loss curves confirmed overfitting, with training loss approaching zero while validation loss plateaued. For future improvement, the neural network architecture is prepared to leverage much larger genomic datasets (e.g., The Cancer Genome Atlas with 10,000+ samples) where deep learning could discover additional differential factors among larger populations.

# Conclusion

This project demonstrated that a multi-model machine learning approach can successfully identify clinically significant gene biomarkers associated with breast cancer survival duration. Three key biomarkers were identified: MAPT, CHEK1, and AURKA, all of which have been independently validated in published cancer genomics research.

The hypothesis was partially supported. As predicted, different models contributed complementary findings: linear regression identified genes with proportional effects on survival, while random forest revealed non-linear gene interactions. The neural network did not perform well due to the small dataset size, as anticipated, but remains prepared for future application with larger datasets.

These biomarkers could enable physicians to personalize breast cancer treatment by assessing individual patients' MAPT, CHEK1, and AURKA expression levels. For example, patients with high CHEK1 expression (associated with poor prognosis) might benefit from more aggressive treatment or CHEK1 inhibitor therapy, while patients with high MAPT expression (associated with better prognosis) might be candidates for less intensive treatment regimens. This personalized approach could improve survival duration and quality of life for breast cancer patients. Unlike existing programs that primarily predict binary survival outcomes (alive vs. deceased), this multi-model approach predicts continuous survival duration and identifies specific gene biomarkers. This provides more actionable clinical information for treatment personalization.

Although the METABRIC dataset provided valuable clinical and gene-expression information, its sample size was still relatively small for deep learning, and some variables contained missing values that required imputation. While median and mode imputation allowed the models to use more patient records, this approach may reduce biological nuance because it estimates missing values rather than measuring them directly. Future studies should use larger and richer datasets, such as TCGA, that include broader genomic, clinical, and treatment-response information (The Cancer Genome Atlas Research Network et al., 2013). In addition, because this project was completed using freely accessible computing resources, model complexity and hyperparameter tuning were limited. Access to greater computational power could allow for more extensive neural network optimization, cross-validation, and survival-analysis methods that account for censoring. These improvements could strengthen both prediction accuracy and biomarker discovery in future versions of the project.

# References

1. Ali, H. R., Dawson, S.-J., Blows, F. M., Provenzano, E., Pharoah, P. D., & Caldas, C. (2012). Aurora kinase A outperforms Ki67 as a prognostic marker in ER-positive breast cancer. *British Journal of Cancer*, 106(11), 1798–1806. https://doi.org/10.1038/bjc.2012.167
2. Al-kaabi, M. M., Alshareeda, A. T., Jerjees, D. A., Muftah, A. A., Green, A. R., Alsubhi, N. H., Nolan, C. C., Chan, S., Cornford, E., Madhusudan, S., Ellis, I. O., & Rakha, E. A. (2015). Checkpoint kinase1 (CHK1) is an important biomarker in breast cancer having a role in chemotherapy response. *British Journal of Cancer*, 112(5), 901–911. https://doi.org/10.1038/bjc.2014.576
3. Bonneau, C., Gurard-Levin, Z. A., Andre, F., Pusztai, L., & Rouzier, R. (2015). Predictive and Prognostic Value of the TauProtein in Breast Cancer. *Anticancer Research*, 35(10),

5179–5184.
4. Callari, M., Sola, M., Magrin, C., Rinaldi, A., Bolis, M., Paganetti, P., Colnaghi, L., & Papin, S. (2023). Cancer-specific association between Tau (MAPT) and cellular pathways, clinical outcome, and drug response. *Scientific Data*, 10(1), 637. https://doi.org/10.1038/s41597-023-02543-y
5. *Cancer Facts & Figures*. (2026). American Cancer Society. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2026/2026-cancer-facts-and-figures.pdf
6. Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 34(2), 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x
7. Fong, Y., Evans, J., Brook, D., Kenkre, J., Jarvis, P., & Gower-Thomas, K. (2015). The Nottingham Prognostic Index: Five- and ten-year data for all-cause Survival within a Screened Population. *The Annals of The Royal College of Surgeons of England*, 97(2), 137–139. https://doi.org/10.1308/003588414X14055925060514
8. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.
9. METABRIC Group, Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Langerød, A., Green, A., … Aparicio, S. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403), 346–352. https://doi.org/10.1038/nature10983
10. Pereira, B., Chin, S.-F., Rueda, O. M., Vollan, H.-K. M., Provenzano, E., Bardwell, H. A., Pugh, M., Jones, L., Russell, R., Sammut, S.-J., Tsui, D. W. Y., Liu, B., Dawson, S.-J., Abraham, J., Northen, H., Peden, J. F., Mukherjee, A., Turashvili, G., Green, A. R., … Caldas, C. (2016). The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nature Communications*, 7(1), 11479. https://doi.org/10.1038/ncomms11479
11. Siegel, R. L., Kratzer, T. B., Wagle, N. S., Sung, H., & Jemal, A. (2026). Cancer statistics, 2026. *CA: A Cancer Journal for Clinicians*, 76(1), e70043. https://doi.org/10.3322/caac.70043
12. The Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., & Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10), 1113–1120. https://doi.org/10.1038/ng.2764
13. Wu, M., Pang, J.-S., Sun, Q., Huang, Y., Hou, J.-Y., Chen, G., Zeng, J.-J., & Feng, Z.-B. (2019). The clinical significance of CHEK1 in breast cancer: A high-throughput data analysis and immunohistochemical study. *International Journal of Clinical and Experimental Pathology*, 12(1), 1–20.