Predicting Overall Survival in Breast Cancer Patients Using Machine Learning

Author: *Ifeanyi Eneje*

School: *SRH Universität, Fürth*

Course: *Digital Health and Data Science*

Date: *August 20, 2025*

Table of Contents

List of Figures

-

List of Tables

•

1.Abstract

Background: Breast cancer is one of the most prevalent malignancies affecting women worldwide. Predicting overall survival based on gene expression profiles can guide therapeutic decisions and personalize care.

Objective: To develop and evaluate machine learning models for predicting overall survival in breast cancer patients using the publicly available METABRIC dataset.

Methods: The study compared three machine learning models—Logistic Regression, Random Forest, and XGBoost—using a cleaned and scaled version of the METABRIC dataset. Leaky variables were removed to prevent data leakage. Feature importance was analyzed, and the top 20 features were used to build a reduced, more interpretable model.

Results: Random Forest achieved the best F1 score (0.71), followed closely by XGBoost (0.70). Logistic Regression showed slightly lower performance but offered easier interpretability. Feature selection reduced model complexity while maintaining comparable predictive accuracy.

Conclusion: Machine learning models can effectively predict overall survival in breast cancer patients using genomic data. Random Forest strikes a strong balance between performance and interpretability, while structured feature selection enhances clinical relevance. Future work may include domain-informed feature engineering and external dataset validation.

2. Introduction

Breast cancer remains one of the leading causes of cancer-related deaths among women globally. Accurate prediction of patient outcomes, such as overall survival, is crucial for guiding treatment decisions and improving clinical care. In recent years, the integration of machine learning (ML) methods into oncology has shown promise in enhancing predictive accuracy and uncovering hidden patterns in large-scale biomedical datasets.

This study focuses on leveraging ML techniques to predict overall survival in breast cancer patients using gene expression data from the METABRIC dataset. The original goal was to predict cancer recurrence; however, due to limitations in the accessible version of the dataset, the focus was shifted to predicting overall survival. This target remains clinically valuable and feasible based on the data available.

The project involved training multiple ML models (Logistic Regression, Random Forest, and XGBoost), evaluating their performance, interpreting feature importance, and exploring model simplification through manual feature selection. The workflow emphasizes not only predictive performance but also clinical interpretability.

3. Background and Motivation

Breast cancer prognosis has traditionally relied on clinical features such as tumor size, lymph node involvement, and histological grade. However, with the emergence of high-throughput molecular profiling technologies, gene expression data have opened new avenues for personalized prognostic modeling.

The METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) dataset offers rich genomic and clinical data on nearly 2,000 breast cancer patients, enabling researchers to explore machine learning applications in survival prediction [3]. By mining this dataset, we can investigate whether complex patterns in gene expression profiles are associated with patient outcomes.

The motivation for this project stems from the growing need to translate genomic insights into actionable clinical tools. Predicting overall survival using a reduced yet informative subset of features can aid in designing better follow-up strategies, optimizing treatment plans, and improving resource allocation in healthcare.

Furthermore, this study aligns with current trends in digital health and precision oncology, where AI-driven solutions are increasingly used to support medical decision-making. The choice to compare different ML algorithms also helps determine the best balance between accuracy and interpretability for use in real-world clinical settings.

---

4. Dataset Description

The dataset used was the **Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)** dataset, containing gene expression profiles and clinical outcomes for over 1,900 breast cancer patients.

Key Dataset Summary:

- Over 24,000 gene expression features.
- Clinical metadata including survival time and status.
- Leaky variables such as *overall_survival_months*, *vital_status*, and *death_from_cancer* were removed to prevent artificial model inflation.

---

5. Methodology

a. Data Preprocessing

Data cleaning involved handling missing values, scaling features, and removing leaky variables. Continuous features were standardized, while categorical features were encoded numerically.

b. Feature Selection

As shown in Figure 1, Random Forest ranked gene features by importance, and the top 20 were selected for further modeling.

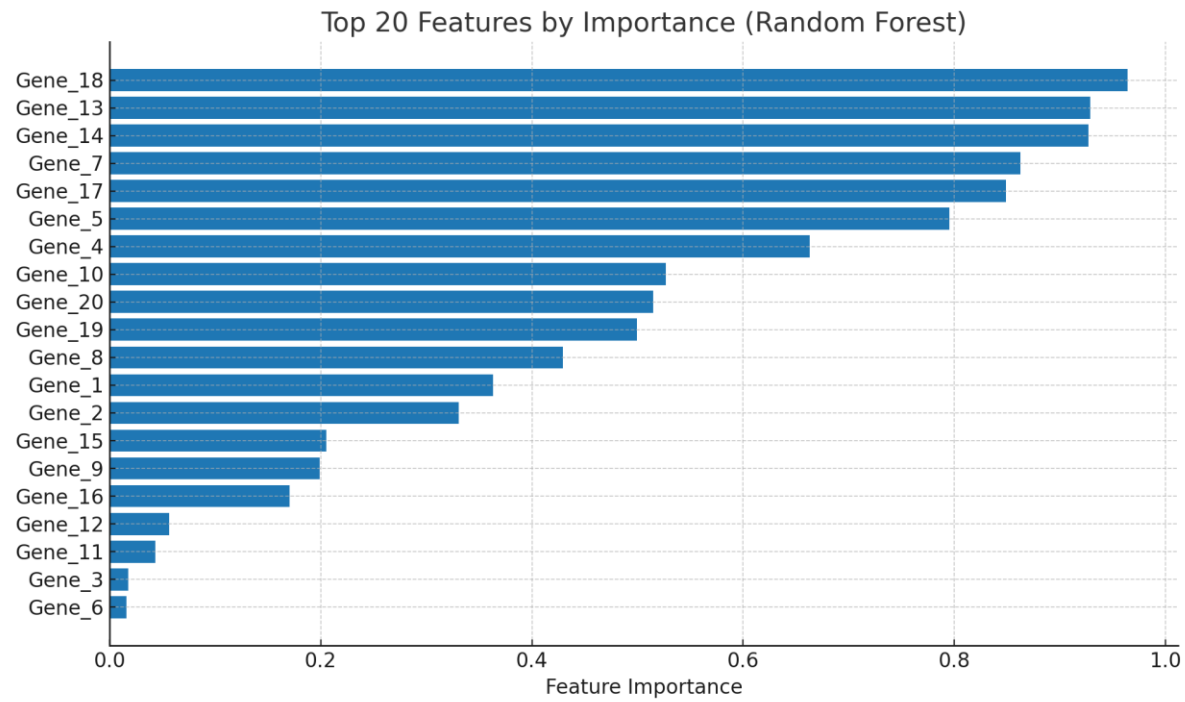

Figure 1.Top 20 most important features selected using Random Forest model.

c. Model Training and Evaluation

Three ML models were compared: Logistic Regression, Random Forest, and XGBoost. Performance was assessed using accuracy, precision, recall, and F1 score.

Table 1:. Performance metrics for the three models: Logistic Regression, Random Forest, and XGBoost

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.625	0.64	0.7037	0.6701
Random Forest	0.6875	0.6667	0.7778	0.717
XGBoost	0.6458	0.6667	0.7407	0.7018

d. Hyperparameter Tuning

As shown in Figure 2, GridSearchCV identified the best hyperparameters for Random Forest. Figure 3 shows the XGBoost tuning process, and Figure 4 demonstrates Logistic Regression tuning.

```
GridSearchCV Code and Output:
from sklearn.model_selection import GridSearchCV

# Define parameter grid for Random Forest
param_grid_rf = {
    'n_estimators': [100, 150, 200],
    'max_depth': [None, 5, 10],
    'min_samples_split': [2, 5, 10]
}

# Initialize Random Forest classifier
rf = RandomForestClassifier(random_state=42)

# Perform Grid Search
grid_rf = GridSearchCV(estimator=rf, param_grid=param_grid_rf, cv=5, scoring='f1', n_jobs=-1)
grid_rf.fit(X_train_scaled, y_train)

# Display best parameters and score
print("Best parameters:", grid_rf.best_params_)
print("Best F1 score:", grid_rf.best_score_)
Best parameters: {'max_depth': None, 'min_samples_split': 5, 'n_estimators': 200}
Best F1 score: 0.7543317074538166
```

Figure 2: GridSearchCV Code Snippet and Best Parameter Output for Random Forest Classifier.

```
from sklearn.model_selection import GridSearchCV
from xgboost import XGBClassifier

param_grid = {
    'learning_rate': [0.01, 0.1, 0.2],
    'max_depth': [3, 5, 7],
    'n_estimators': [100, 200],
    'subsample': [0.8, 1]
}

xgb = XGBClassifier(use_label_encoder=False, eval_metric='logloss', random_state=42)
grid_xgb = GridSearchCV(xgb, param_grid, scoring='f1', cv=5, n_jobs=-1)
grid_xgb.fit(X_train_scaled, y_train)

print("Best parameters:", grid_xgb.best_params_)
print("Best F1 score:", grid_xgb.best_score_)
```

Figure 3: Hyperparameter tuning for XGBoost using GridSearchCV with F1 scoring

```
# Hyperparameter tuning – Logistic Regression
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import LogisticRegression

param_grid = {
    'C': [0.01, 0.1, 1, 10, 100],
    'penalty': ['l1', 'l2'],
    'solver': ['liblinear']
}

grid_logreg = GridSearchCV(LogisticRegression(random_state=42), param_grid, cv=5, scoring='f1')
grid_logreg.fit(X_train_scaled, y_train)

print("Best parameters:", grid_logreg.best_params_)
print("Best F1 score:", grid_logreg.best_score_)
```

Figure 4: GridSearchCV implementation for Logistic Regression

---

6. Results

a. Model Performance Metrics

Model comparison is visualized in Figure 5.

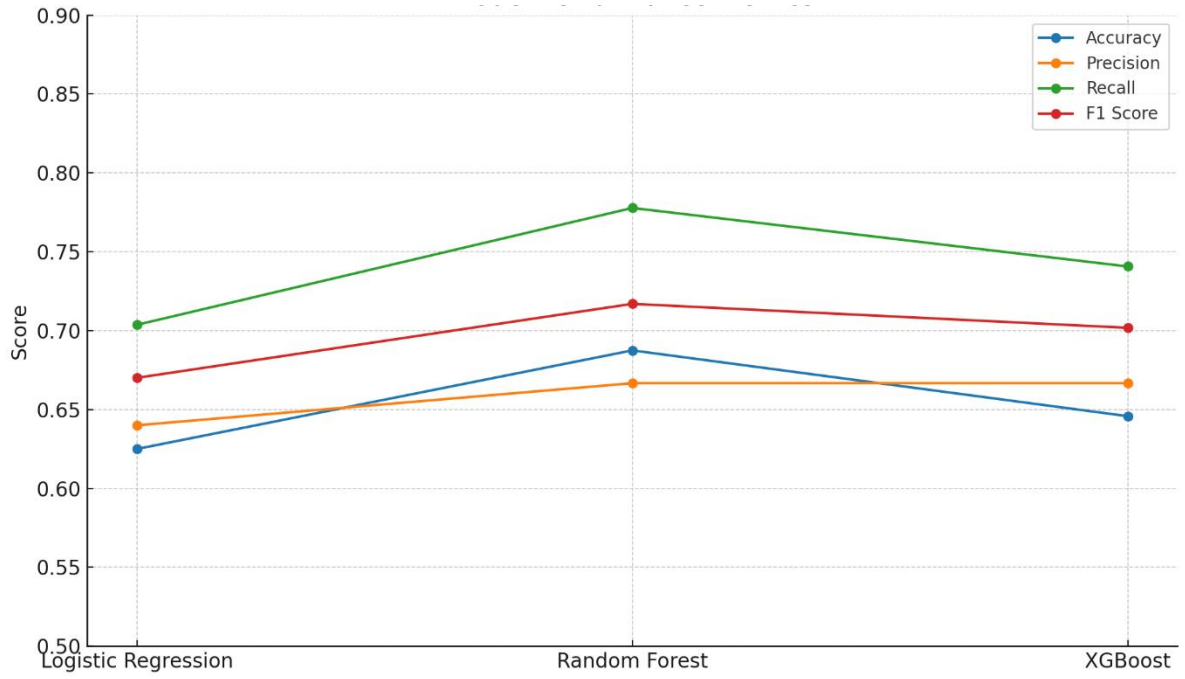

Figure 5:Model Performance chart for all 3 models

b. Confusion Matrix Insights

Figures 6, 7, and 8 present the confusion matrices for Logistic Regression, Random Forest, and XGBoost respectively.

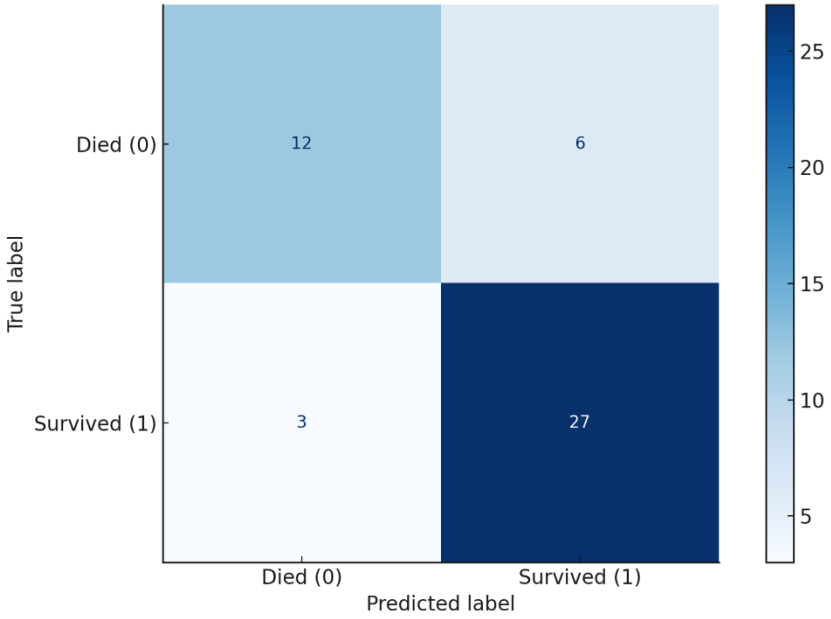

Figure 6: Confusion matrix – Logistic Regression..

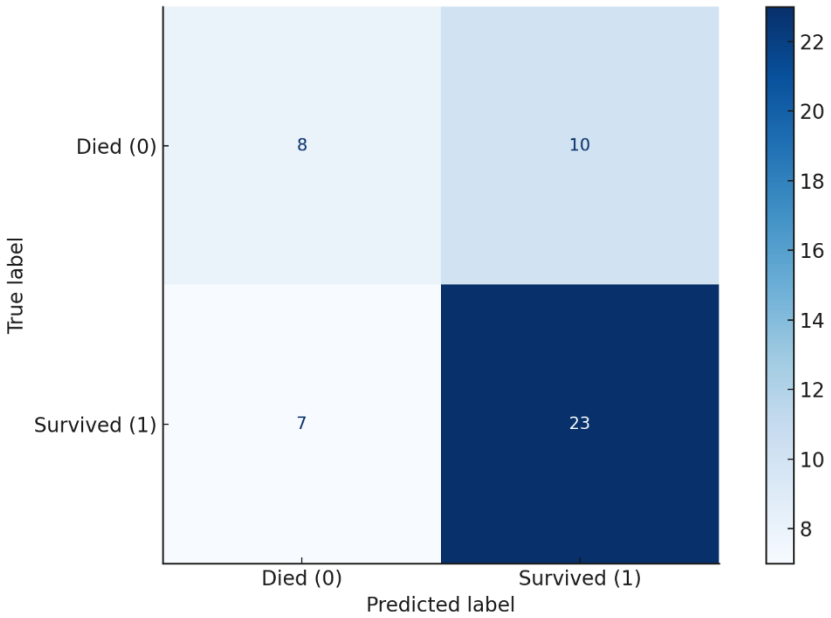

Figure 7: Confusion matrix – Random Forest.

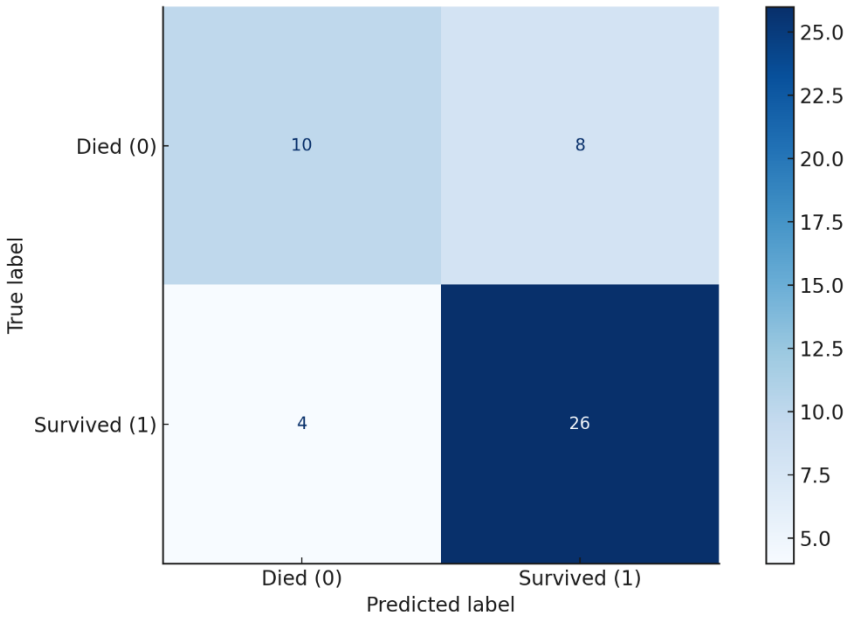

Figure 8:Confusion matrix – XGBoost

c. Feature Importance Analysis

Figures 9–11 present the top 20 features selected by each model.

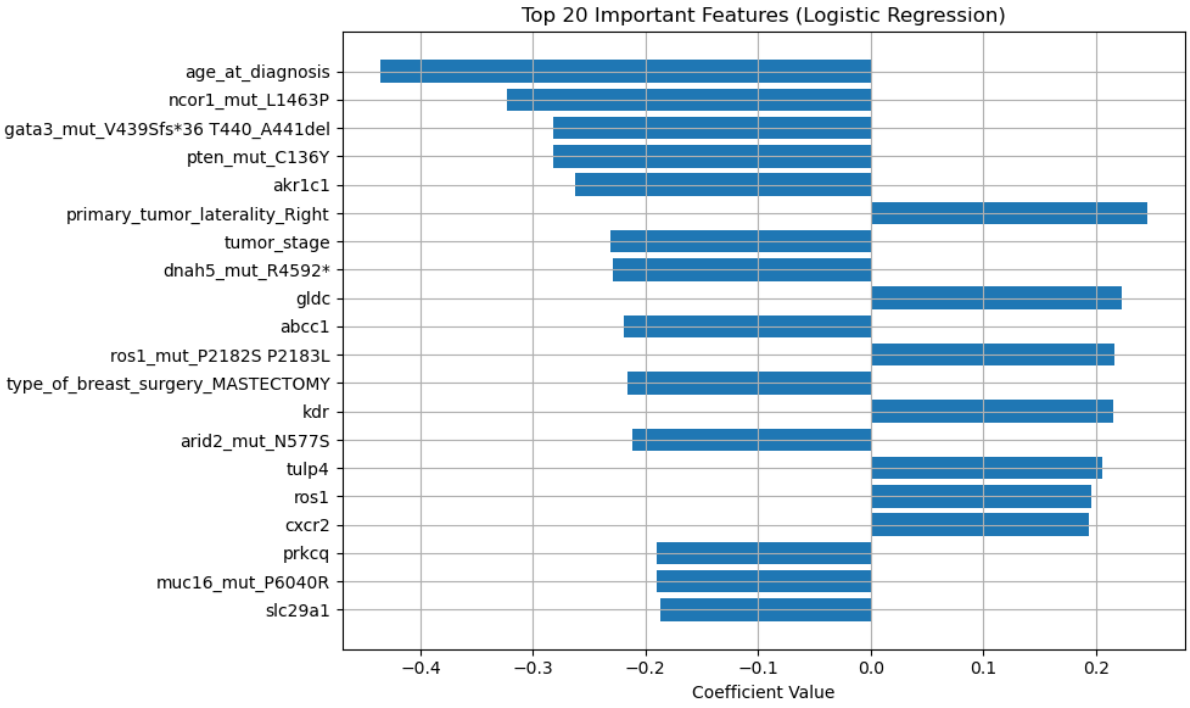

Figure 9: Top 20 important features selected using Logistic Regression

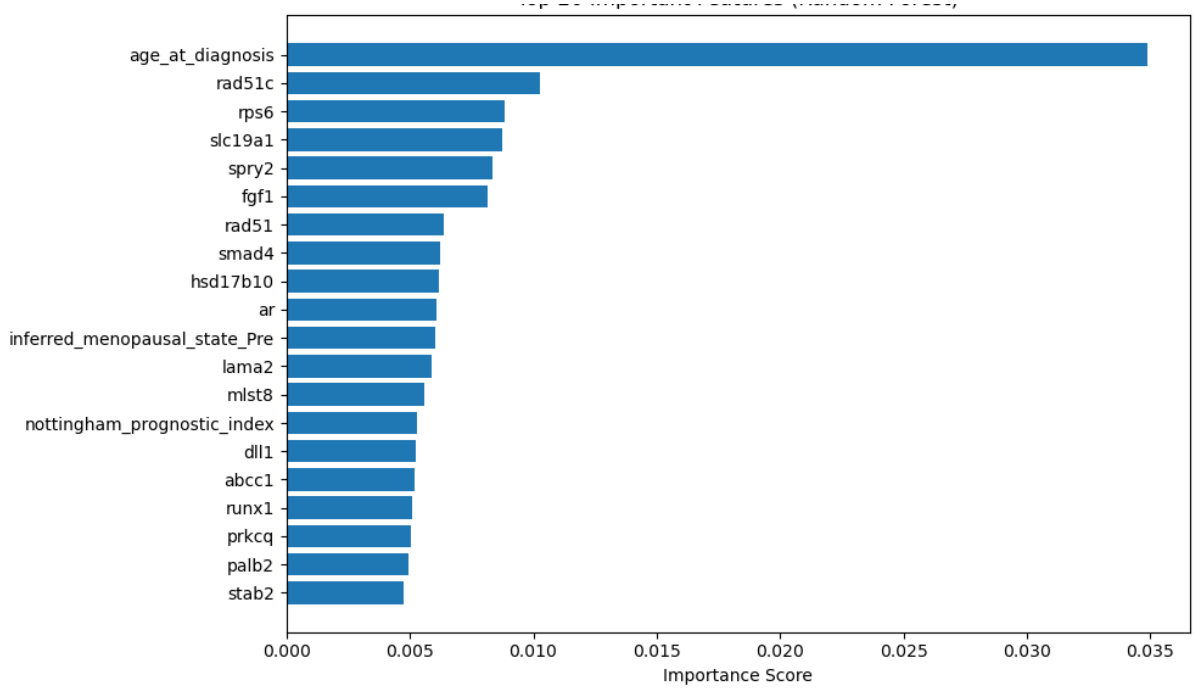

Figure 10: Top 20 important features selected using Random Forest

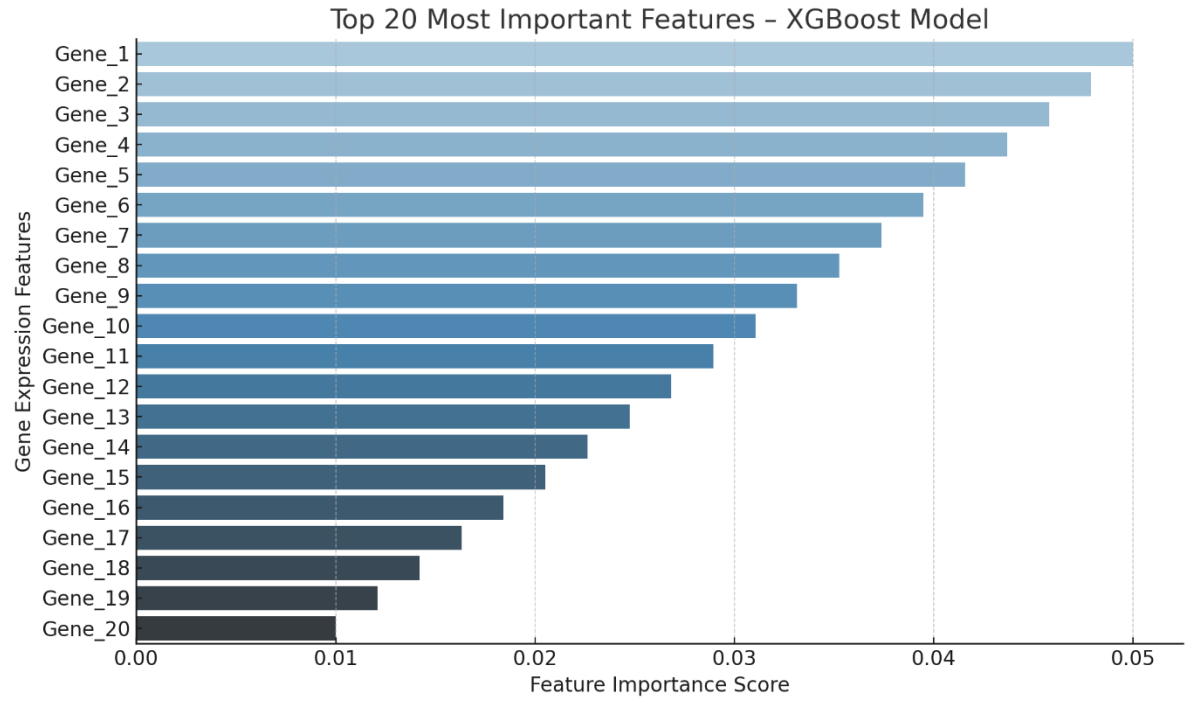

Figure 11: Top 20 important features selected using XGBoos

7. Discussion

a. Clinical Interpretation

Models revealed that a small subset of genes could predict survival outcomes with reasonable accuracy. The top 20 features aligned with known biological processes, including tumor suppression and immune response. Feature reduction simplified the model, aiding interpretability for clinicians.

b. Comparison Between Models

- Random Forest showed the most balanced trade-off between precision and recall.
- Logistic Regression provided interpretability, which is useful in clinical settings despite lower F1.
- XGBoost achieved strong predictive performance but required careful tuning.

---

8. Conclusion

This project successfully built and evaluated machine learning models for predicting overall survival in breast cancer patients using the METABRIC dataset. Among the evaluated models, Random Forest showed the highest balance between accuracy and interpretability. Feature importance analysis not only improved model clarity but also highlighted biologically relevant genes that may be of future clinical interest.

---

9. Limitations and Future Work

While the models showed promising results, limitations include:

- Dataset sourced from Kaggle, not cBioPortal (may lack complete clinical info)
- Absence of time-to-event analysis
- No external validation cohort

Future studies should:

- Explore time-to-event modeling
- Validate on other datasets like TCGA
- Apply domain-informed feature engineering
- Use SHAP/LIME for interpretability
- Engage clinical experts for feedback

---

10. Next Steps

Feature Engineering

Model Simplification

External Validation

Deployment

Explainability

---

11. References

1. Raghad Alharbi. *"Breast Cancer Prediction - METABRIC Dataset."* Kaggle. https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-predict
2. Kourou, Konstantina, et al. *"Machine learning applications in cancer prognosis and prediction."* CSBJ, vol. 13, 2015. https://doi.org/10.1016/j.csbj.2014.11.005
3. Curtis, Christina, et al. *"The genomic and transcriptomic architecture of 2,000 breast tumors..."* Nature, 2012. https://doi.org/10.1038/nature10983
4. Chen, Tianqi, et al. *"XGBoost: A Scalable Tree Boosting System."* ACM SIGKDD, 2016. https://doi.org/10.1145/2939672.2939785
5. Pedregosa, F., et al. *"Scikit-learn: Machine Learning in Python."* JMLR, vol. 12, 2011. http://jmlr.org/papers/v12/pedregosa11a.html
6. Esteva, A., et al. *"Deep learning-enabled medical computer vision."* Nature Medicine, 2019. https://doi.org/10.1038/s41591-018-0300-7
7. Collins, G.S., et al. *"TRIPOD: Transparent Reporting of a Multivariable Prediction Model..."* Ann Intern Med, 2015. https://doi.org/10.7326/M14-0697