Breast Cancer Prognosis: Leveraging Cancer Registry Data for Survival Prediction

CST 489/499

Nicholas Marolla, Federico Murrieta, Rahim Siddiq, Harris Popal

Dr. Alireza Abdoli

Summer 2025

# Executive Summary

The goal of this project is to develop a predictive model for breast cancer survival using clinical data from the SEER cancer registry, sourced from Kaggle. The project focuses on equipping participants with skills in data cleaning, feature selection, and survival analysis using machine learning techniques.

Breast cancer is a leading cause of cancer-related deaths, making early identification of patients at higher risk crucial for improving survival rates. This project aims to predict the likelihood of survival based on clinical features such as age, tumor size, lymph node status, and hormone receptor status. By doing so, it offers a tool that can help clinicians better prioritize high-risk patients and optimize treatment strategies.

The key objectives are: (1) to build a robust model using machine learning methods like survival XGBoost and survival random forest, and (2) to interpret model findings using techniques like survSHAP to identify important prognostic factors. The dataset includes 4,024 patient records with features related to demographics and tumor characteristics. Unlike traditional classification models, survival prediction focuses on estimating the time to event, which is more complex but crucial for patient prognosis.

The anticipated outcomes include the creation of a reliable model for survival prediction and a deeper understanding of key factors influencing breast cancer prognosis. By the end of the project, participants will gain hands-on experience with survival analysis techniques and be able to apply them to real-world health data, contributing to improved decision-making in breast cancer treatment.

In summary, this project aims to advance the use of machine learning in clinical prognosis, providing insights that can improve survival rates and aid in the development of personalized treatment plans for breast cancer patients.

# Table of Contents

[**Executive Summary 2**](#_5ny0aqrjd8sj)

[**Table of Contents 4**](#_cak2uq6mr3e4)

[**Introduction/Background 5**](#_z5nbsg5cx6ty)

[Project Name and Description 5](#_aa7m3nvp26yv)

[Problem and/or Issue in Technology 5](#_m5xu0xhzjqa7)

[Solution to the Problem and/or Issue in Technology 6](#_gqwz52tzpcax)

[**Environmental Scan / Literature Review 6**](#_u22sh6pvzaad)

[**Stakeholders 7**](#_yx2ts520k7dk)

[Ethical Considerations 8](#_ukedc83p9pfv)

[Legal Considerations 8](#_40yejz26en20)

[**Project Goals and Objectives: 9**](#_kesdmojhjvtz)

[**Final Deliverables 10**](#_1fbtgoy8d6tp)

[**Approach/Methodology 10**](#_cuz34iuboa3k)

[**Timeline/Resources 11**](#_hosfe08aa06h)

[**Platform 12**](#_58isn1g1cucv)

[**Risks and Dependencies 13**](#_ekwao2396ex9)

[Risks 13](#_m05oqtn4tsib)

[Dependencies 13](#_a0535kz03jbq)

[**Testing Plan 14**](#_5mfaffdgfuf)

[**Division of Responsibility 14**](#_vzl4nwxhvqbp)

[**References 15**](#_foozx2yxrg4e)

# Introduction/Background

## Project Name and Description

The project, *Breast Cancer Prognosis: Leveraging Cancer Registry Data for Survival Prediction*, aims to build a predictive model using clinical data from the SEER cancer registry. The project utilizes machine learning algorithms to predict the survival rates of breast cancer patients based on various clinical features, such as age, tumor size, lymph node status, and hormone receptor status. This project is for healthcare professionals and data scientists, providing them with a tool to assist in early risk identification and better treatment planning for patients. By improving the accuracy of breast cancer prognosis, this project is crucial for increasing survival rates and optimizing healthcare outcomes.

## **Problem and/or** I**ssue in** T**echnology**

Breast cancer is one of the leading causes of cancer-related deaths globally. Despite advancements in treatment, predicting the survival of patients remains a complex and challenging task. Traditional models focus on binary outcomes (survival or death), but survival prediction requires estimating the time to an event, which is more nuanced and requires advanced analytical techniques. Current models often lack the accuracy needed for personalized treatment planning, leading to potential delays in high-risk patient interventions. Thus, a more accurate, data-driven approach is needed to predict survival rates and provide personalized care strategies.

## Solution to the Problem and/or Issue in Technology

This project will address the problem of inadequate survival prediction by developing a machine learning model that leverages clinical data to predict breast cancer survival probabilities. Using algorithms like XGBoost and random forests, the model will predict the time to an event based on a patient's clinical features. The model will be trained, validated, and tested on a dataset of over 4,000 breast cancer patients, and the findings will be interpreted using methods like survSHAP to identify key prognostic factors. This solution will provide healthcare professionals with actionable insights to improve patient outcomes.

# Environmental Scan / Literature Review

Breast cancer survival prediction is a well-researched area, with several studies focusing on the development of prognostic models. One such study by Phung et al. (2019) reviews various models for breast cancer prognosis, emphasizing the importance of incorporating clinical data into predictive tools. Their systematic review of prognostic models highlights the potential of machine learning techniques to improve prediction accuracy. Similarly, a study by Efthimiou et al. (2024) discusses the application of machine learning in clinical prediction models, offering valuable insights into the development of accurate survival models for different cancers.

Machine learning in survival analysis has gained popularity due to its ability to handle complex datasets and predict outcomes over time. Techniques like XGBoost and survival random forests have been used to develop highly accurate survival prediction models, as demonstrated in works by Krzyziński et al. (2023), who explored the use of these algorithms for time-dependent survival models. This project builds upon these methods, applying them specifically to breast cancer survival prediction using clinical data from the SEER registry.

The availability of large-scale datasets like the SEER cancer registry further strengthens the feasibility of this project. With over 4,000 patient records, the dataset offers a rich resource for training robust models. The use of such data is supported by resources like the PREDICT breast cancer website (Grootes et al., 2024), which provides tools and insights for clinical prediction, illustrating the impact of accurate prognostic models in real-world applications.

# Stakeholders

The stakeholders in this project include healthcare professionals, patients, data scientists, and developers. Healthcare professionals stand to benefit the most from the project as they will gain a predictive tool that assists in decision-making, particularly for high-risk patients. This model will improve early risk identification, enabling more personalized treatment plans and better resource allocation. Patients will benefit indirectly through improved treatment outcomes and survival rates.

Data scientists and developers working on the project will gain valuable experience in applying machine learning techniques to health data and survival analysis, strengthening their skills in data modeling and interpretation. The clients, if any, such as healthcare institutions or research organizations, will benefit from the use of a predictive tool that can contribute to more accurate cancer prognosis and improved patient care.

## Ethical Considerations

Ethical considerations in this project primarily focus on the use of patient data. Since the dataset includes sensitive health information, privacy and confidentiality are critical issues. Strict adherence to data privacy laws, including HIPAA, will be essential in ensuring that patient information is handled responsibly. Steps will be taken to anonymize the data and ensure that the results of the model cannot be traced back to individual patients.

Accessibility is another ethical concern. The tool should be designed to be accessible to healthcare providers with varying levels of technological expertise, ensuring that the model’s predictions can be easily interpreted and applied in clinical settings. Regular audits will be conducted to ensure fairness and transparency in the model’s decision-making process, avoiding any potential bias based on demographic factors.

## **Legal Considerations**

Legal considerations will primarily involve ensuring that all data used in the project is properly sourced and that any necessary permissions are obtained for using the SEER dataset. Compliance with data protection laws, such as GDPR or HIPAA, is crucial to prevent misuse of sensitive health data. Additionally, any intellectual property generated from the project, such as the predictive model or code, will need to be clearly attributed, and any copyrights or licensing requirements for software tools used will be followed.

# **Project Goals and Objectives:**

| **Goals** | **Objectives** |
| --- | --- |
| Develop a robust survival prediction model for breast cancer. | Train a model using survival XGBoost and random forests. |
| Improve the accuracy of breast cancer survival predictions. | Validate and test the model using SEER dataset. |
| Provide interpretable insights into key prognostic factors. | Use survSHAP to explain the model's feature importance. |

| **Goals** |
| --- |
| Develop a robust survival prediction model for breast cancer. |
| Improve the accuracy of breast cancer survival predictions. |
| Provide interpretable insights into key prognostic factors. |

## 

| **Objectives** |
| --- |
| Train a model using survival XGBoost and random forest. |
| Validate and test the model using the SEER dataset. |
| Use survSHAP to explain the model's feature importance. |

# Final Deliverables

The final deliverables for this project will include the developed predictive model, a detailed report on the model’s performance and findings, and a tool for healthcare professionals to use in predicting breast cancer survival. Additionally, a GitHub repository will be provided with the project’s code, and a user guide will be available to assist healthcare providers in applying the model. The project will be complete upon receiving client approval and demonstrating the model's effectiveness in real-world applications.

# Approach/Methodology

Our approach to developing the breast cancer survival prediction model will be systematic and iterative, drawing on an Agile framework to manage progress toward each milestone. The project will commence with data acquisition, specifically utilizing the SEER cancer registry data sourced from Kaggle. Key initial processes include thorough data preprocessing: cleaning the data, appropriate encoding and transformation of clinical features such as age, tumor size, lymph node status, and hormone receptor status. Regular weekly team meetings will facilitate task allocation and progress tracking, ensuring alignment with our project goals.

Following data preparation, the core of our methodology will focus on the machine learning pipeline. This includes feature engineering and selection to identify the most significant predictors of survival from the clinical data. We will then train and validate models. The performance of these models will be rigorously evaluated using the SEER dataset and appropriate survival analysis metrics to achieve our goal of improving prediction accuracy. Finally, to provide interpretable insights into key prognostic factors, we will employ techniques like survSHAP, aligning with our objective to make the model's findings understandable and actionable for healthcare professionals.

# Timeline/Resources

| Phase | Tasks | Duration |
| --- | --- | --- |
| Project Initialization & Setup | Finalize proposal. Set up project environment. Detailed review of the dataset and documentation. | 1 Week |
| Data Preprocessing | Verify the dataset from Kaggle. Clean data, eliminate outliers, and address censored data. | 1 Week |
| Feature Engineering | Research and identify new features for feature engineering. Apply the selected features and finalize for modeling. | 1 Weeks |
| Model Development & Training | Begin tuning hyperparameters for selected models, cross-validate, and train | 2 Weeks |
| Model Evaluation | Evaluate model performances, compare the performance of various models. Adjust feature engineering to optimize performance. | 1 Weeks |
| Front-End | Develop the interface for users. Finish project report and prepare code repo and documentation. | 1 Week |
| Final Report & Presentation | Finishing touches for all deliverables | 1 Week |

# **Platform**

This project utilizes Streamlit and Google Collab as the primary platforms for model development, deployment, and user interaction. Streamlit has been chosen for building the interactive web-based application due to its simplicity, rapid development capabilities, and intuitive interface. As an open source Python library, Streamlit allows for the integration of machine learning models, enabling the team to create a user-friendly application. Additionally, the platform’s ability to provide real-time interactivity means users can explore different model configurations, view prediction results instantly, and gain insights into the key prognostic factors identified by the model.

For model development, Google Collab is the preferred platform due to its cloud-based environment, which provides free access to GPU and TPU resources, which makes it ideal for handling computationally intensive tasks such as training survival prediction models. Google Colab also supports the full Python ecosystem, allowing for the use of libraries like Random Forest, XGBoost, and survival analysis tools, which are essential for this project. Its collaborative features also make it easy for team members to share code, monitor progress, and ensure version control.

# Risks and Dependencies

What are the risks that can affect the completion of the project on time or at all? What are the dependencies for completing the project? Relate these to your project objectives and milestones.

## **Risks**

The dataset from Kaggle might contain unforeseen levels of missing data, inconsistencies, errors, or biases not apparent. This may significantly extend the data preprocessing phase and affect the features for modeling, thus impacting its performance in a live environment. Additionally, the chosen models may not achieve the desired level of accuracy and generalizability. This would affect the time for feature and parameter tuning, significantly increasing the time spent on optimizing the model to achieve satisfactory results.

## **Dependencies**

This project’s results heavily depend on the beginning phase of data preprocessing. A high-quality dataset is crucial for the success of achieving desirable results in proper breast cancer prognosis.

# Testing Plan

The primary focus of the testing plan for this project is to ensure high accuracy and reliability of the survival prediction models, which is quite important for providing meaningful insights in a clinical context. The performance of the machine learning models will be evaluated using established survival analysis metrics, including the Concordance Index (C-Index), which measures the model’s ability to correctly rank patient survival times, and the Brier Score, which quantifies the accuracy of the probabilistic predictions over time. Moreover, continuous monitoring of these metrics during model training will ensure the final model achieves optimal predictive accuracy.

# Division of Responsibility

Nicholas Marolla. Federico Murrieta. Rahim Siddiq. Harris Popal.

# References

Efthimiou, O., Seo, M., Chalkou, K., Debray, T., Egger, M., & Salanti, G. (2024). Developing

clinical prediction models: A step-by-step guide. *BMJ*, 386, 1–10.

https://doi.org/10.1136/bmj.2024.065367

Grootes, I., Wishart, G. C., & Pharoah, P. D. (2024). An updated PREDICT breast cancer

prognostic model including the benefits and harms of radiotherapy. *NPJ Breast Cancer, 10*(1), 6. https://doi.org/10.1038/s41523-024-00612-1

Krzyziński, M., Spytek, M., Baniecki, H., & Biecek, P. (2023). SurvSHAP (t): Time-dependent

explanations of machine learning survival models. *Knowledge-Based Systems, 262*, 110234. https://doi.org/10.1016/j.knosys.2023.110234

Phung, M. T., Tin Tin, S., & Elwood, J. M. (2019). Prognostic models for breast cancer: A

systematic review. *BMC Cancer, 19*, 1–18. https://doi.org/10.1186/s12885-019-5894-2

Spooner, A., Nguyen, B., & McKinney, S. (2020). A comparison of machine learning methods

for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports, 10*(1), 20410. https://doi.org/10.1038/s41598-020-77481-0