Beyond the Cure: Modeling the Risk of Breast Cancer Recurrence

Brandon Thomas

bthomas@bellarmine.edu

16 January 2025

Executive Summary

The goal of this project is to develop a predictive model for breast cancer recurrence, leveraging advanced machine learning techniques to improve clinical decision-making and patient outcomes. Breast cancer recurrence, whether local or metastatic, remains a significant challenge in cancer care, and accurate prediction of recurrence risks is crucial for personalized treatment planning and long-term monitoring. The project will utilize a comprehensive dataset, which includes clinical and demographic features like tumor size, lymph node status, and age.

To tackle this, Python will serve as the core programming language, utilizing a suite of specialized libraries. Pandas and NumPy will be used for data cleaning, preprocessing, and numerical computations. scikit-learn will handle traditional machine learning models such as logistic regression, decision trees, and SVMs. For more robust, non-linear modeling, XGBoost and Random Forest will be employed, taking advantage of ensemble learning and capturing intricate interactions between features. TensorFlow/Keras will be used to construct and train deep neural networks, offering the potential to capture complex patterns in the data.

For model evaluation and interpretation, SHAP will provide insights into feature importance, while Matplotlib and Seaborn will be used for static visualizations, and Plotly/Dash will create interactive dashboards. Finally, Jupyter Notebook will facilitate experimentation, and Git/GitHub will manage version control. This comprehensive approach aims to provide a more accurate, interpretable, and scalable model for predicting breast cancer recurrence.

Project Idea

This project aims to develop an advanced predictive model for breast cancer recurrence using machine learning techniques. By leveraging a dataset that includes clinical and demographic features, the goal is to create a model that accurately predicts the risk of recurrence, enabling personalized treatment plans and long-term patient monitoring. The project will utilize a variety of machine learning algorithms, including traditional models like logistic regression and decision trees, as well as more advanced techniques such as XGBoost, Random Forest, Support Vector Machines (SVMs), and deep learning with TensorFlow/Keras to attempt to capture complex relationships in the data.

Background

Despite significant advances in the detection and treatment of breast cancer, recurrence remains a persistent challenge for patients and clinicians. Breast cancer recurrence, whether local or metastatic, affects approximately 20-30% of patients within five years post-treatment, depending on the initial stage and subtype of the disease (Early Breast Cancer Trialists' Collaborative Group, 2021). Predicting the likelihood of recurrence can provide critical insights for personalized treatment planning and long-term monitoring, which are essential for improving survival rates and quality of life. However, the complexity of recurrence patterns, influenced by tumor biology, treatment responses, and patient-specific factors, necessitates robust predictive models capable of handling potentially non-linear and multifactorial relationships.

Breast cancer recurrence presents a major obstacle in long-term cancer care. Recurrence may manifest as local relapse or as metastatic disease affecting distant organs, significantly impacting prognosis and survival. Key challenges include inaccurate predictions, limited personalization, and underutilization of existing data. Existing clinical tools for predicting recurrence often fail to achieve optimal sensitivity and specificity. For example, tools such as the Nottingham Prognostic Index, while widely used, rely on linear assumptions and may not capture the nuanced interactions between variables like tumor grade, size, and lymph node involvement (Gao et al., 2022). Conventional approaches rely heavily on population-level statistics and do not provide individualized risk assessments using patient specific measures. This can result in either overtreatment or undertreatment, with adverse consequences for patient outcomes. Advances in data collection from electronic health records (EHRs), genomic studies, and imaging technologies have produced rich datasets. However, traditional statistical methods often fail to capitalize on the full predictive potential of these complex and high-dimensional data sources.

Various tools and methods have been developed as existing solutions to address the problem of breast cancer recurrence prediction. Clinical guidelines-based models like Adjuvant! Online and the Nottingham Prognostic Index estimate recurrence risks using clinical and pathological features. While useful as general guides, these models are rigid and lack the adaptability to incorporate emerging data such as molecular subtypes or treatment innovations (Blows et al., 2010). Machine learning (ML) approaches, including logistic regression, decision trees, and support vector machines (SVM), have shown promise in improving prediction accuracy. For instance, studies have demonstrated that tree-based methods like Random Forests can outperform traditional statistical models in some contexts (Zhang et al., 2020). However, these methods often struggle with

generalizability, particularly when applied to diverse patient populations, an ensuing challenge in medical research. Neural networks and deep learning models, capable of capturing complex, non-linear patterns, have emerged as cutting-edge tools in recurrence prediction. For example, convolutional neural networks (CNNs) have been used to analyze histopathological images, while recurrent neural networks (RNNs) have shown promise in temporal data analysis. Despite their potential, these models require significant computational resources and risk overfitting when applied to small datasets (Yasaka et al., 2018).

The breast cancer dataset I obtained from Kaggle contains features such as tumor size, lymph node status, and patient demographics, offering a valuable resource for model development. By leveraging this dataset, we can address the limitations of existing methods and develop predictive models that combine traditional clinical insights with the power of ML and deep learning techniques.

Modeling

Predicting breast cancer recurrence is a complex task that requires models capable of capturing intricate patterns in clinical and biological data. Four machine learning models—XGBoost, Random Forest, Support Vector Machines (SVMs), and Neural Networks—stand out for their suitability to address this challenge. XGBoost (Extreme Gradient Boosting) is a highly efficient and scalable algorithm that excels at capturing non-linear interactions and hierarchical relationships between features. Its built-in regularization mechanisms help reduce overfitting, while its ability to handle missing data and noise makes it particularly effective for clinical datasets. Similarly, Random Forest offers robust performance by leveraging ensemble learning, combining predictions from multiple decision trees to reduce variance and improve generalization. Its ability to provide feature importance scores makes it highly interpretable, which is crucial for understanding the factors contributing to recurrence risk.

Support Vector Machines (SVMs) are particularly well-suited for datasets with overlapping classes or high-dimensional feature spaces, as they maximize the margin between classes and effectively model non-linear relationships through kernel functions. This makes SVMs valuable for distinguishing borderline cases, such as patients with intermediate recurrence risk. On the other hand, Neural Networks (NNs) are powerful tools for uncovering complex, non-linear relationships in diverse datasets. Their adaptability allows for specialized architectures, such as deep neural networks for tabular data, convolutional neural networks for imaging, or recurrent

neural networks for temporal patterns. While neural networks require more data and computational resources, they excel in scenarios involving diverse and high-dimensional features.

The choice of model often depends on the dataset's characteristics. For smaller datasets like the one I am using, Random Forest and SVMs are preferred due to their robustness and lower data requirements. For larger or more complex datasets, XGBoost and Neural Networks offer greater flexibility and predictive power. Additionally, interpretability is a critical consideration; while XGBoost and Random Forest provide insights into feature importance, Neural Networks require advanced techniques like SHAP or LIME for explainability. Together, these models provide a comprehensive toolkit for addressing the challenge of breast cancer recurrence prediction, each offering unique strengths to accommodate varying data and clinical needs.

Tools

Python will serve as the core programming language, leveraging a suite of powerful libraries. For data manipulation, Pandas and NumPy will be used for cleaning and numerical operations, while scikit-learn handles initial machine learning models and preprocessing. XGBoost, Random Forest, Support Vector Machines (SVM), and TensorFlow/Keras will be utilized for building and training models, with a focus on efficiency and non-linear relationships. Visualization will be facilitated by Matplotlib, Seaborn, and Plotly/Dash for static and dynamic charts, while SHAP (SHapley Additive exPlanations) will provide model interpretability. Jupyter Notebook will be used for development and documentation, with version control managed through Git/GitHub. Additional tool libraries like Imbalanced-learn will address class imbalance, and Hyperparameter Tuning libraries like Optuna will optimize model performance.

Conclusion

In conclusion, this project represents a comprehensive approach to addressing the complex challenge of predicting breast cancer recurrence through the integration of advanced machine learning techniques and clinical insights. By utilizing the dataset available from Kaggle, which includes key clinical and demographic features, and leveraging cutting-edge algorithms such as XGBoost, Random Forest, Support Vector Machines, and Neural Networks, this project aims to develop accurate and personalized predictive models.

The use of Python and its versatile libraries, including Pandas, NumPy, and scikit-learn, ensures robust data preprocessing and exploration, while visualization tools like Matplotlib, Seaborn, and Plotly/Dash facilitate clear

communication of insights. To ensure interpretability and clinical relevance, SHAP will provide detailed feature attribution, making model predictions more transparent and applicable.

Ultimately, this project strives to advance breast cancer care by overcoming the limitations of traditional predictive tools. By creating interpretable and scalable models, it aims to empower clinicians with more precise recurrence risk assessments, aiding in personalized treatment planning and long-term monitoring. This innovative approach has the potential to improve patient outcomes and contribute to the broader goal of enhancing survivorship and quality of life for breast cancer patients.

References

Blows, F. M., Driver, K. E., Schmidt, M. K., et al. (2010). Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. PLoS medicine, 7(5), e1000279.

Early Breast Cancer Trialists' Collaborative Group (EBCTCG). (2021). Long-term outcomes after radiotherapy or surgery for early-stage breast cancer. The Lancet Oncology.

Gao, Z., Zhang, Y., Styblo, T. M., et al. (2022). Machine learning applications in predicting breast cancer recurrence and treatment response. Cancer Informatics, 21, 117693512211073.

Lichman, M.: UCI machine learning repository, University of California, School of Information and Computer Science, Irvine, CA (2019). http://archive.ics.uci.edu/ml/datasets/breast+cancer

Yasaka, K., Akai, H., Kunimatsu, A., et al. (2018). Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. Radiology, 286(3), 887-896.

Zhang, X., Yang, P., & Xu, Z. (2020). Applications of machine learning in breast cancer diagnosis and prognosis. Cancer Biology & Medicine, 17(4), 943.