Early Detection and Diagnostic Insights

CSBP 4502 – Group 7

Carolina Perez
University of Colorado Boulder
Boulder CO USA
cape5274@colorado.edu

Cody Folgmann
University of Colorado Boulder
Boulder CO USA
cody.folgmann@colorado.edu

Dain Kim
University of Colorado Boulder
Boulder CO USA
dain.kim@colorado.edu

Problem Statement

Cancer survival outcomes have been a large focus of medical research, particularly the field of early detection/diagnosis technologies that offer the potential for improving patient prognosis, as well as clinical and genetic data to find patterns in cancer diagnosis. Despite advances in diagnostic methods, predicting patient survival rates based on early detection remains a significant challenge. This project aims to bridge this gap by investigating how early detection factors such as tumor size, biopsy results, and patient demographics can be linked to improved survival rates and outcomes. By identifying and analyzing these key variables, we seek to construct a predictive model that will assess cancer survival probabilities.

The knowledge we plan to apply in this research comes from various fields, including data analysis, machine learning, and healthcare studies/research. We intend to utilize machine learning techniques to process and analyze our clinical dataset that includes factors such as tumor size and biopsy results and additional clinical data to identify the most significant factors influencing survival rates. In addition to clinical variables, we will also consider patient demographics such as age, gender, and overall health history, as these can have a profound impact on survival. By integrating all these data points, we aim to create a robust model that not only predicts survival probability but also provides insights into the relationship between early detection and patient survival and create understanding how early detection factors, such as tumor size and biopsy results, correlate with improved survival, as early interventions based on these factors could lead to better treatment strategies and patient education.

With increased accurate survival predictions, healthcare providers could prioritize treatment for high-risk patients, optimize healthcare resources and improve survival rates. This project could reveal whether certain demographic factors, when integrated with clinical data, can further refine survival predictions. This research will show aspects of early detection are the most influential in determining patient survival. By identifying these, the study could inform future diagnostic practices and patient care strategies, leading to more personalized and effective treatments. By better understanding the key factors that contribute to survival, this project could help healthcare professionals identify at-risk individuals sooner, leading to timely intervention that can improve long-term outcomes. The goal is to develop a model that will not only predict survival probabilities but also help shape future diagnostic and treatment guidelines, resulting in improved patient outcomes and more effective cancer care and patient education.

Previous work/ Literature survey

There is an extensive list of models involving diverse types of cancer and cancer detection methods:

1. *Breast Cancer Survival Prediction Model*

This model focuses on predicting survival rates for breast cancer patients by analyzing clinical and genetic data. Its goal was to assess the likelihood of survival based on factors such as: tumor size, genetic

mutations, and overall health history of the patient and is aimed at helping medical professionals make informed decisions about treatment plans, diagnosis, and patient education on early detection. It could be a helpful tool towards patients because it offers a clear picture of their prognosis and helps create a bridge between genetic and clinical data.

2. *Lung Cancer Survival Prediction Model*

This model was developed in 2021 by Y Wu, and it emphasizes the importance of early detection for improving survival rates in lung cancer patients. Factors in this data include Age, smoking history and tumor characteristics are used to predict survival rate for these patients. The main goal of this model is to emphasize early detection and screening for patients and is aims for those patients who are at elevated risk for this – those with a smoking history or family history and could be a great resource for patient education.

3. *Socio-Economic Factors and Cancer Survival*

This model considers socio–economic factors and their impact on cancer survival rates such as: patient age, income, education, and healthcare access. This model is used to show the correlation between the patient socio-economic factors and their chance of survival and provides a different viewpoint of how these factors impact chance of survival compared to the usual factors different studies and models show – clinical and genetic data. This model could be used when discussing public health and policy decisions to improve health care and treatment options and accessibility with cancer patients of diverse backgrounds

Data Preprocessing and Modeling

To begin, we will preprocess the data by handling missing values—using mean imputation for numerical fields (like tumor size) and mode imputation for categorical fields (such as biopsy results). Any duplicates in the dataset will be removed to ensure data integrity. We will standardize the numerical features using Z-score normalization, and for categorical variables, such as tumor type, we will apply label encoding to make them machine-readable. Due to the large size of the dataset, feature selection will play a crucial role. We will employ Recursive Feature Elimination (RFE) to identify the most relevant predictors, including tumor size, genetic risk, and patient age, while removing redundant features.

For modeling, we will use a Random Forest Classifier. This model is well-suited for structured medical data and provides built-in feature importance scores, helping us interpret the results. The data will be split into an 80% training set and a 20% testing set. We will use 5-fold cross-validation to improve the model's generalization and prevent overfitting. Random Forest models are known to be effective for survival prediction tasks because they can handle complex, non-linear relationships in medical data.

To evaluate performance, we will focus on the F1-score, which balances precision (the number of correctly predicted survivors) and recall (how well the model identifies actual survivors). Additionally, we will use the ROC-AUC score to measure how well the model distinguishes between survival and non-survival cases. This approach aims to create a robust and interpretable model that can shed light on the key survival factors in brain tumor prognosis.

Data Set

URL:
https://www.kaggle.com/datasets/ankushpanday1/brain-tumor-prediction-dataset

For this project we are using the Brain Tumor Prediction Dataset, containing approximately 250,000 data points and 22 attributes. These attributes include clinical data such as tumor size, genetic risks, patient symptoms, and some socio-economic details such as patient lifestyle.

This dataset provides a realistic representation of medical data that can be used for predictive modeling and survival analysis. It includes key information about tumor location, growth rate, and survival rate, offering comprehensive insights into the factors that influence brain tumor progression and patient outcomes. Additionally, it allows for diverse and inclusive analysis of how different patient

demographics and environmental factors impact tumor survival. The data's scale and scope make it particularly suitable for classification tasks and detailed survival analysis models.

Evaluation Methods

Cross-validation will be used to test the predictive performance of our model. The primary method for this will be K-fold cross validation. K-fold cross validation partitions the dataset into K subsets, then retains one subset while using the remaining K-1 subsets to train the model. The resulting model can then be tested against the retained subset. This process is repeated K times using each subset as the test set exactly once, and then the results are averaged together, letting us use all the data efficiently for both training and testing.

K-fold cross-validation is an effective method for evaluating model performance as it ensures efficient data utilization by allowing each data point to be used for both training and testing while reducing overfitting risk. It is particularly useful for small datasets, providing a more reliable performance estimate compared to a single train-test split. A small K-value (e.g., K=2) may lead to high bias, while a large K-value (e.g., K=20) can increase variance. A K-value of 10 is recommended as a starting point. An improperly chosen K-value can significantly skew the outcomes and is the main risk of this evaluation method.

For medical datasets where class imbalance is a concern, K-fold cross-validation should be complemented with additional metrics such as ROC-AUC, precision-recall curves, and F1-score to provide a comprehensive evaluation. ROC-AUC assesses the model's ability to distinguish between classes by plotting the true positive rate (sensitivity) against the false positive rate, with a higher AUC indicating better discrimination (Bradley, 1997). However, in highly imbalanced datasets, precision-recall curves may provide a more informative evaluation by focusing on the trade-off between precision and recall rather than overall classification thresholds (Saito & Rehmsmeier, 2015). If computational resources allow, nested cross-validation, which combines K-fold with an inner validation loop for hyperparameter tuning, can offer even more robust performance estimation (Varma & Simon, 2006).

Tools

GitHub will be used for version control, collaboration, and documentation, with a well-maintained README to detail the dataset, project structure, and analysis steps. The repository will house all project code, including data processing with Pandas and NumPy, machine learning models using Scikit-learn and TensorFlow/PyTorch and visualizations created with Matplotlib and Seaborn. This approach ensures smooth collaboration, easy access to resources, and effective tracking of progress.

Milestones – Completion Dates

1. Data Preprocessing – Mar 14
2. Feature Selection – Mar 21
3. Model Implementation – Apr 04
4. Model Evaluation – Apr 18
5. Final Model Selection & Report Writing – Apr 25

References:

[1] Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7), 1145-1159.

[2] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432.

[3] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432.

[4] Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, 7(1), 91.