# Early Detection and Diagnostic Insights

CU Boulder – CSPB 4502
Spring 2025 Group 7

Carolina Perez (cape5274@colorado.edu)
Cody Folgmann (cody.folgmann@colorado.edu)
Dain Kim (dain.kim@colorado.edu)

# Agenda

- Introduction
- Related Work
- Data Set
- Techniques Used
- Results
- Applications
- Visualization

# Introduction

- This project aims to develop a machine learning model that integrates clinical data (tumor size, biopsy results) and patient demographics to predict cancer survival rates and identify key factors influencing outcomes, particularly related to early detection. .
- By enhancing prediction accuracy and understanding the impact of early interventions, the study seeks to inform diagnostic practices, optimize treatment strategies, and improve patient education and care. The attributes that we care about the most are:
  - Turmor size
  - Genetic risk

# Dataset

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

- ~250,000 datapoints
- 22 attributes
- Demographic as well as clinical information on patients
- Patient outcomes as well as initial conditions

# Techniques Used - Model

- Random Forest Classifier
  - Attributes have high likelihood of non-linear correlations
  - Accepts multiple variable types (categorical, continuous, etc)
  - Provides attribute rankings, helping identify key predictors of survival and minimizing impact of noisy features

# Techniques Used – Validation

- K-Fold Cross-Validation
  - K = 5 selected after initial testing, ran 5 times
  - minimizes risk of a particular training set driving the model
- Confusion matrix
  - Visualizes true/false positives and negatives
- Classification Report
  - Summarizes precision, recall, and F1 score

# Results

- **Blood pressure, genetic risk, and age emerged as the most predictive variables**, highlighting the value of using accessible patient data to estimate survival rates early.
- **While tumor size showed a strong link to survival, its utility is limited by the need for clinical diagnosis**, reinforcing our goal to prioritize non-invasive predictors for earlier intervention.
- **Symptom severity and growth rate had unexpectedly low importance**, which supports our objective to reassess traditional assumptions and focus on data-driven indicators in predictive modeling.

# Applications

Early Risk Identification and Integration into Routine Care

Resource Prioritzation

Public Awareness and Prevention

# Visualization

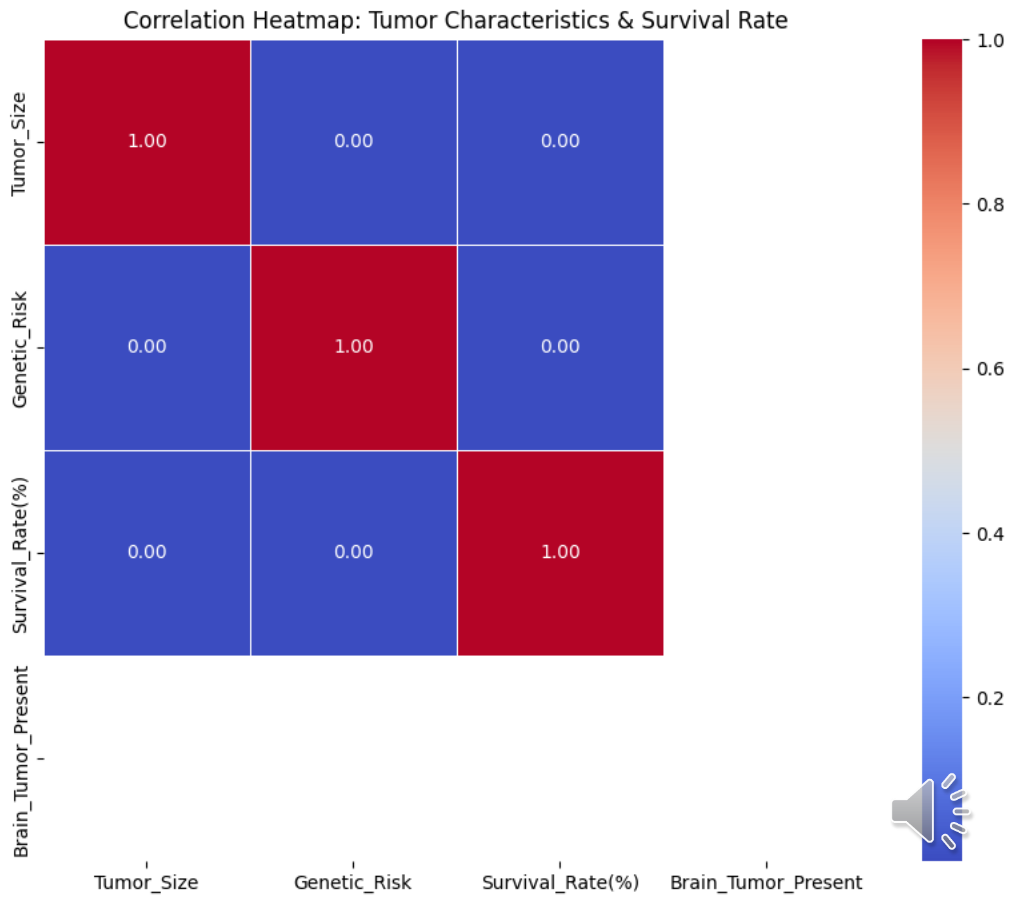
- In this project, we used data visualisation techniques to investigate correlations between features and understand relationships within a dataset.

- Visual insights can help suggest where to improve or change models or preprocessing steps.

- By highlighting hidden patterns and feature behaviour, visualisation can guide you to make better modelling decisions and strengthen your predictive performance.

# Visualization – Correlation Heatmap

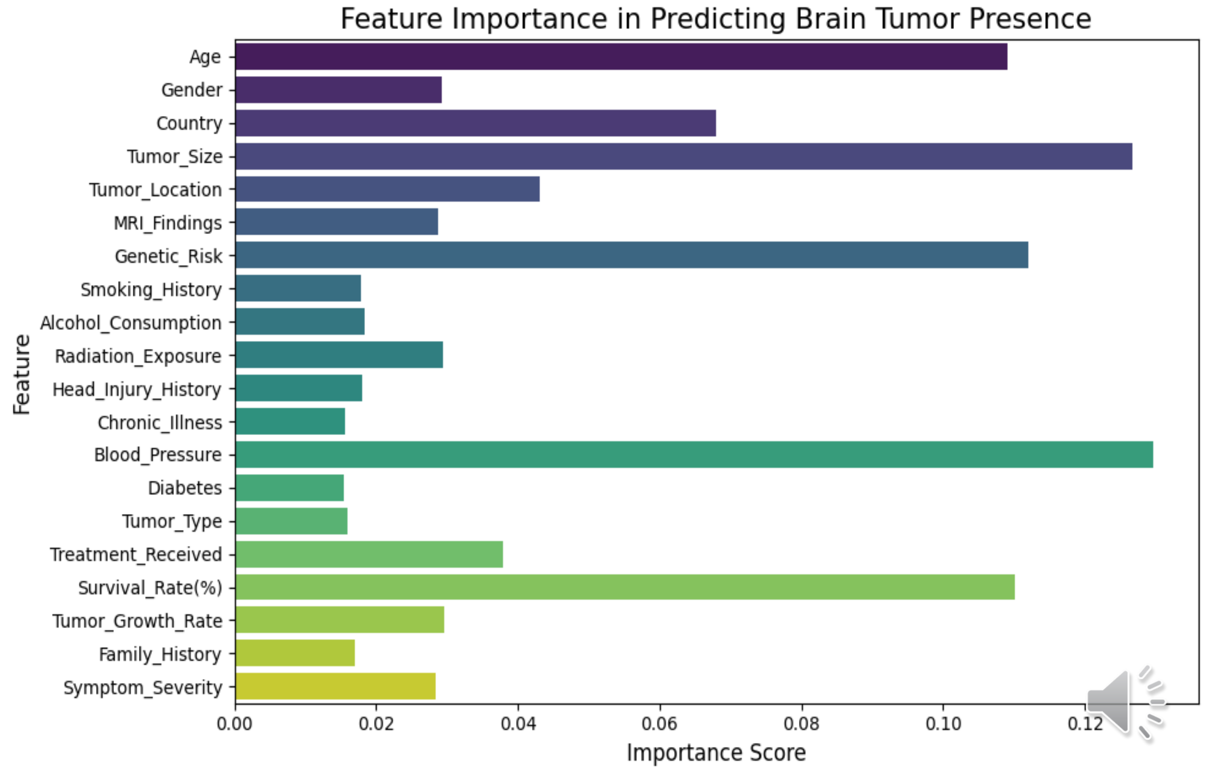
- It shows how different features are related to each other and target(Brain_Tumor_Present).
- From here, we assume that higher Tumor_Size might negatively correlate with Survival_Rate(%).



# Visualization – Feature Importance

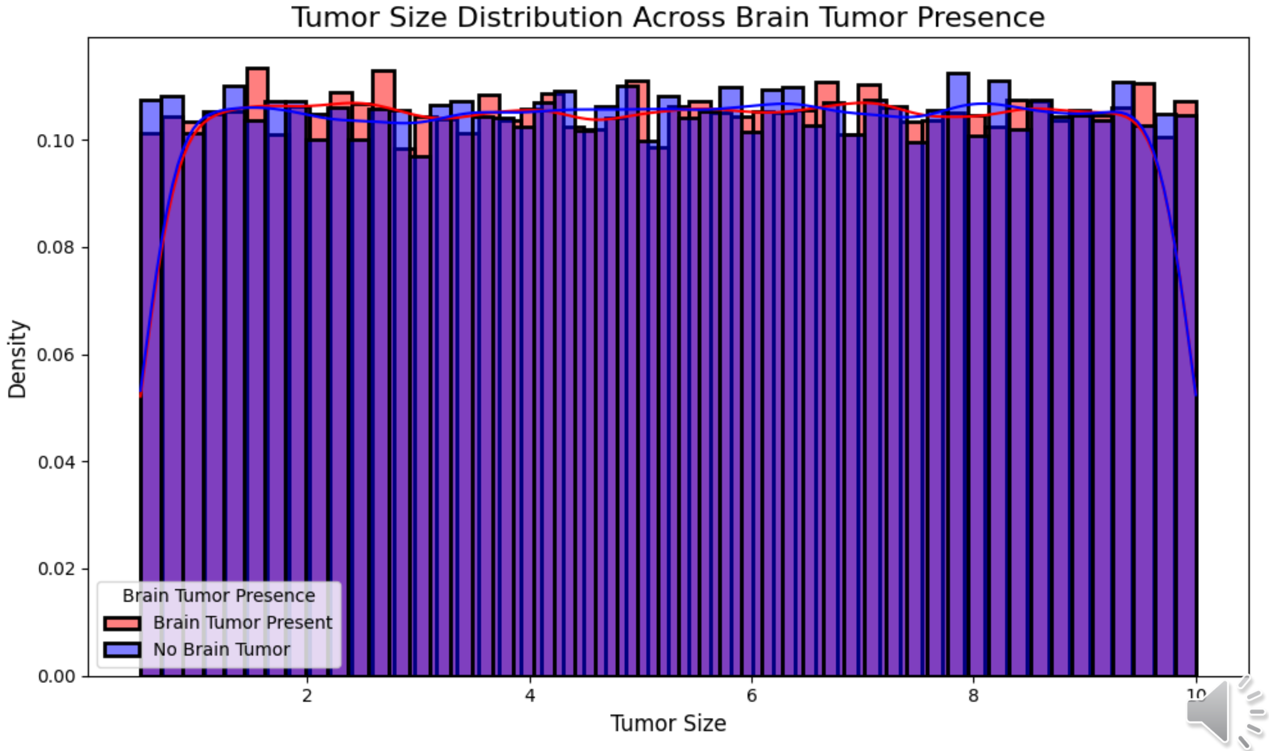
- Random Forest evaluates how much each feature improves classification through decision trees, capturing both linear and nonlinear relationships.

- It helps in understanding which patient attributes most affect brain tumor presence -> Genetic_Risk and Tumor_Size could be most influential.



# Visualization – Tumor size distribution

- Histogram shows the distribution of tumor sizes for patients with and without a brain tumor.
- It compares how tumor size patterns across the two class and check if the tumor size thresholds affected patients or not -> patients with tumors tend to have larger tumor sizes on average.



# Modelling

- High accuracy and ability to handling complex data -> **A random forest classifier** for brain tumor prediction
- Its built using multiple decision trees for robust prediction and set random processes  for data sampling and feature selection ensuring the model training and results are reproducible across different runs.

# Evaluation

- **Train-Test Split**

  Splitting the data into a training set (to build the model) and a separate testing set (to evaluate its performance on unseen data) to avoid overfitting

- **Cross-Validation**

  Training and testing the model multiple times on different subsets.

  -> Provided more stable measure from 5 folds

# Conclusion

- We built a machine learning model (a random forest classifier) to predict the presence of brain tumours, using visualisations such as correlation heatmaps to understand key feature relationships.

- The model, which is particularly sensitive to tumour size and genetic risk, demonstrated robust and consistent predictive performance through cross-validation, suggesting high confidence in future predictions.