AIC-201: Predicting 5 Year Mortality in Colorectal Cancer Patients using a ANN

By: Hardik (Josh)

1 Introduction

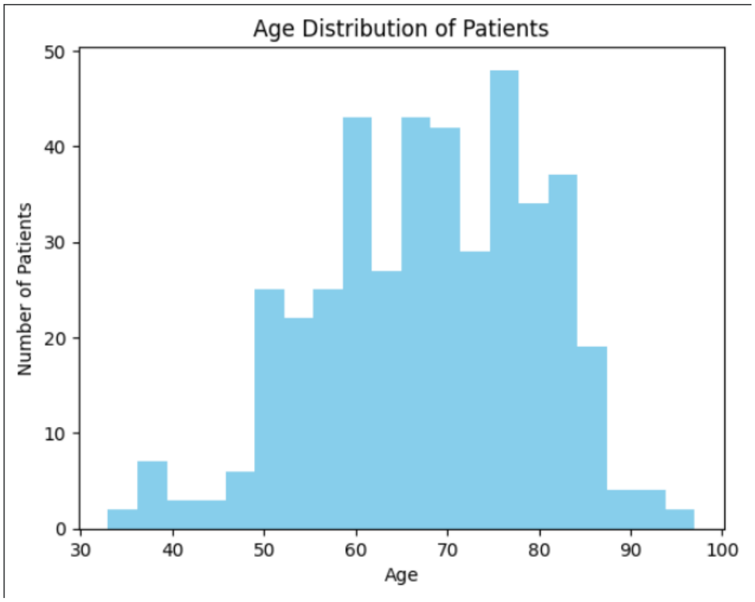
The goal for this project was to design, train and evaluate an Artificial Neural Network (ANN) to predict the survival outcome of colorectal cancer in five years using the SR386 cohort from the SurGen dataset. The dataset contains clinical, demographic and genetic information of 427 patients, with the target variable indicating whether a patient died within five years of diagnosis. The task is clinically relevant because early and accurate mortality prediction can support oncologists in patient risk assessment, treatment planning and post surgery care strategies.

ANN's are suitable for this task because they can model complex, non-linear relationships between clinical, demographic, and genetic features and the survival outcome.

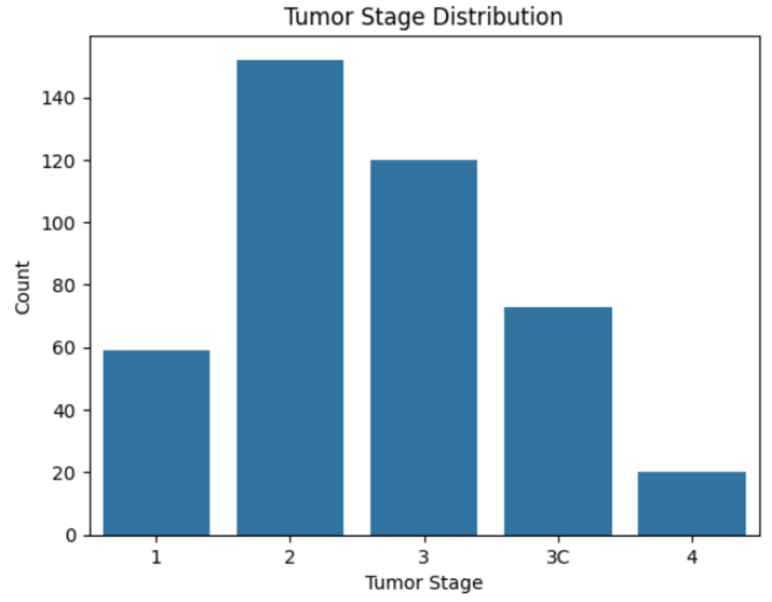
2 Data Exploration and Analysis (EDA)

The dataset was first explored to understand the structure, missing values and distribution of features. Key exploratory steps included viewing the data shape, inspecting missing values and visualizing important features such as age, tumor stage, sex and survival outcomes.

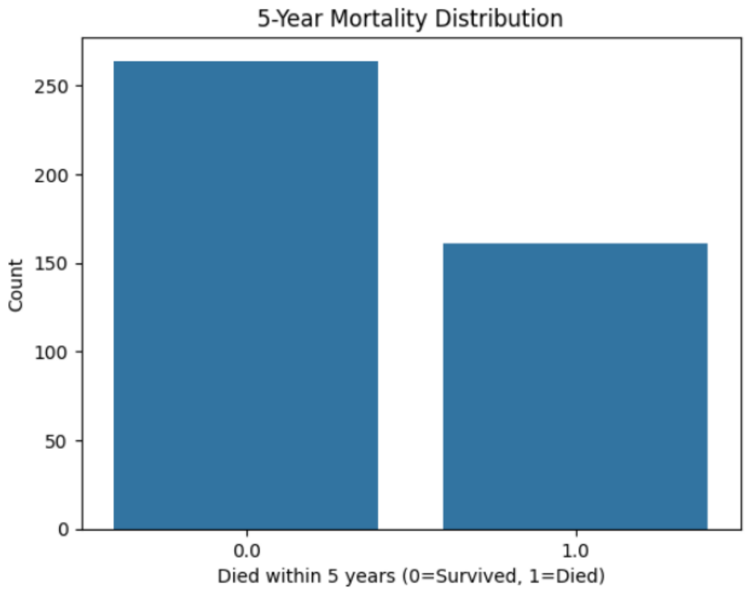
Visualizations showed that most patients were aged somewhere between 55 and 75 years old, suggesting that age could be important in survival chances. The class distribution of the target variable 'died_within_5-years' was slightly imbalanced, with a higher proportion of survivors. Categorical columns such as tumor stage, MSI status and sex were explored to understand the relationship with the target variable.



[Figure 1. Age distribution histogram]



[Figure 2. Tumor stage plots]



[Figure 3. Target variable distribution (died_within_5_years)]

3 Data cleaning and Preprocessing

Missing numerical values were imputed using the mean while categorical values were imputed using the most frequent category. Columns with over 30% missing data or very low variability were dropped to ensure model quality.

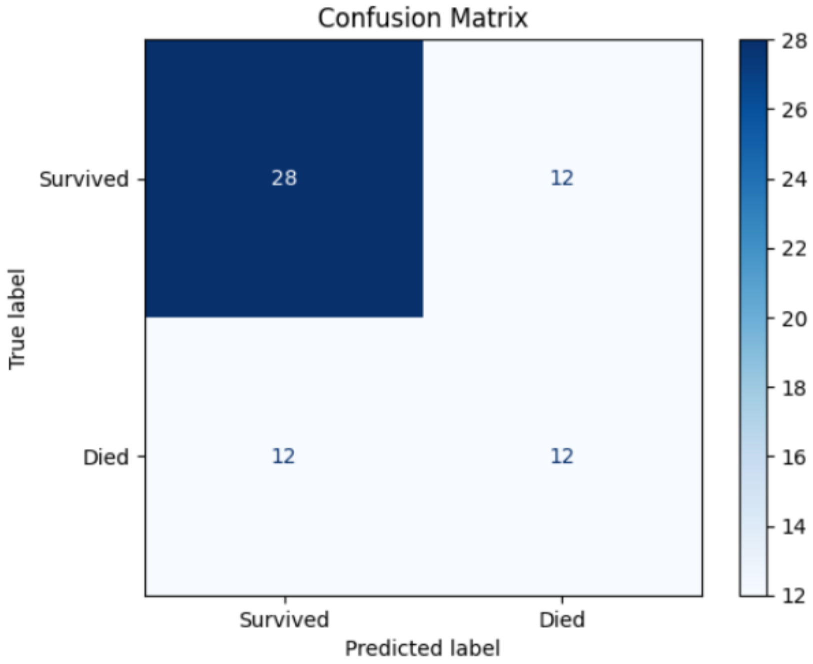
Categorical features were one-hot-encoded, and numerical features were standardized using StandardScaler to ensure consistent scaling.

The dataset was split into **70:15:15**, **Training**, **Validating** and **Testing** respectively using stratified sampling to preserve class proportions across all splits.

4 Baseline Model

The baseline ANN consisted of two dense layers with 64 and 32 neurons both using ReLU activations, and a dropout layer (0.3) to reduce overfitting. The output layer used a sigmoid activation for binary classification. The model was trained for 30 epochs using the Adam optimizer and binary cross-entropy loss.

The confusion matrix showed that the model correctly predicted mist survivals and deaths, but there were still false positives (predicting death for a survivor) and false negatives (missing a true death). In a medical context, false negatives are more serious, as they can lead to undertreated or delayed care.



[Figure 4. Confusion Matrix for Baseline Model]

5 Model Improvement

Several improvement techniques were used to enhance performance:

SMOTE (Synthetic Minority Oversampling Technique): Balanced the dataset by generating synthetic examples of the minority class.

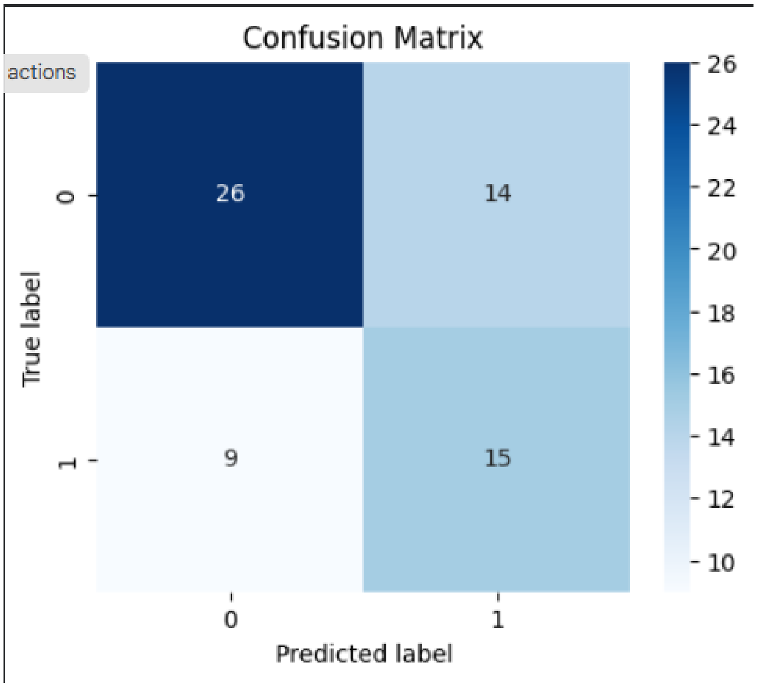
Class Weights: Increased the penalty for misclassifying minority-class samples.

Regularization (L2): Reduced overfitting by adding a weight penalty.

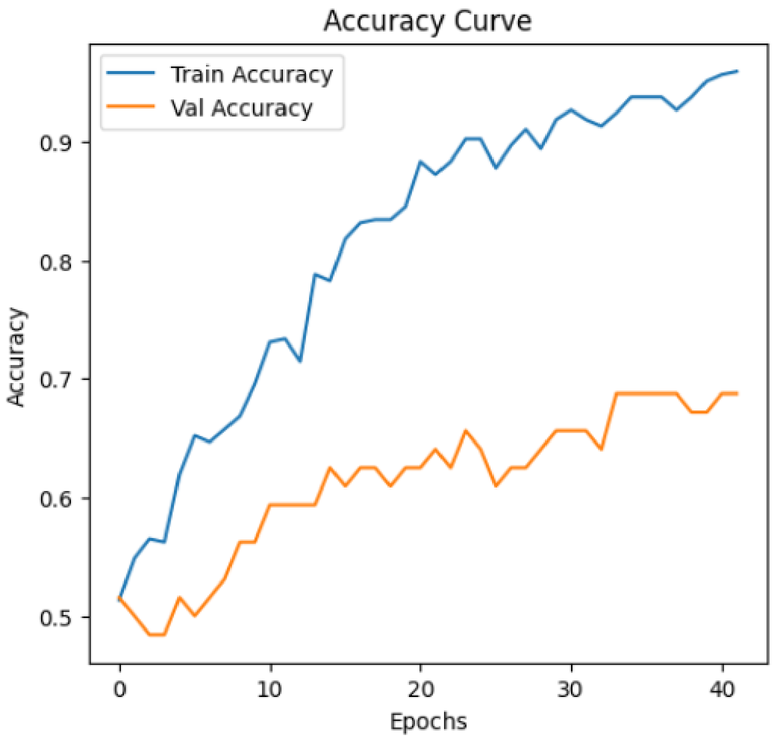
Dropout (0.4): Randomly dropped neurons during training to improve generalization.

Learning Rate Tuning (0.0005): Combined with EarlyStopping and ReduceLROnPlateau callbacks for more stable convergence.

After applying these techniques, the model showed smoother learning curves and slightly improved recall. This suggests it became more effective at identifying patients who did not survive, which is critical in clinical predictions.



[Figure 5. Confusion Matrix for Improved Model]



[Figure 6. Training and Validation Curves]

6 Discussion

The improved ANN demonstrated a balanced trade off between precision and recall. Since the task involves predicting mortality, recall is the more important metric, as missing patient who is at risk of death (false negatives) could lead to inadequate medical attention.

While accuracy remained moderate, the model's stability improved after class balancing and regularization.

7 Conclusions

This project successfully built and evaluated ANN to predict five year mortality among colorectal cancer patients, through thorough preprocessing, model design and optimization, the final model achieved balanced predictive performance and valuable insights into future relevance.

Although not yet suitable for clinical deployment, this work demonstrates how AI can support healthcare decision-making by identifying high risk patients and enabling earlier, more targeted interventions.