Application of Machine Learning Models in Predicting Cancer Patient Survival: A Comparative Analysis Using Logistic Regression, AdaBoost, SVM, and Random Forest

M. Azadmarzabadi
Department of Physics Faculty of Science, Arak University, Arak 384817758, Iran
M. H. Mousavi
Department of Physics Education Farhangian University isfahan, Iran

Abstract

This study investigates the application of machine learning algorithms to predict patient survival outcomes using a synthetic cancer dataset. A comprehensive data preprocessing pipeline, including advanced feature engineering and selection techniques, was employed to maximize data quality and relevance. Four classification algorithms were evaluated, with ensemble methods such as AdaBoost and Random Forest significantly outperforming traditional approaches due to their ability to model complex, non-linear interactions. Rigorous hyperparameter optimization and cross-validation ensured the robustness and generalizability of the models, addressing potential overfitting and enhancing predictive reliability. The findings emphasize the critical role of data preprocessing, model selection, and systematic evaluation in developing effective machine learning solutions for healthcare. This work demonstrates the transformative potential of machine learning in predictive analytics, offering valuable insights into its deployment for personalized medicine and evidence-based clinical decision-making. By showcasing the synergy between advanced algorithms and high-quality data, this study contributes to advancing machine learning applications in healthcare.

Keywords: **Machine Learning, Cancer Survival Prediction, Feature Importance, Ensemble Learning, models comparison**.

Introduction

Advancements in machine learning have revolutionized the field of healthcare, offering powerful tools for analyzing complex datasets and predicting patient outcomes. In oncology, machine learning methods are increasingly used to predict survival, assess treatment effectiveness, and identify critical prognostic factors. Accurate prediction of survival outcomes can assist clinicians in developing personalized treatment plans, improving patient care, and optimizing resource allocation. However, applying machine learning to healthcare data requires careful consideration of data preprocessing, feature selection, and model evaluation to ensure reliable results.

This study focuses on the application of machine learning techniques to predict survival outcomes in a synthetic cancer dataset. By analyzing features such as age, tumor size, lymph node involvement, cancer stage, and treatment type, we explore the capabilities of four widely used machine learning models: Logistic Regression, AdaBoost, Support Vector Machine (SVM), and Random Forest. These models were chosen to represent a range of approaches, from interpretable linear methods to robust ensemble techniques.

The goals of this study are threefold: (1) to evaluate the predictive performance of each model in terms of accuracy and recall, (2) to analyze the importance of clinical features in predicting survival outcomes, and (3) to identify the most effective algorithm for this dataset. This work not only highlights the potential of machine learning in cancer survival prediction but also underscores the importance of proper data preprocessing and model evaluation for achieving accurate and actionable insights. [5]

Data and Methods

Dataset Overview

The dataset utilized in this study is a synthetic cancer dataset that includes a variety of clinical and demographic features, such as **Patient Age**, **Tumor Size**, **Lymph Node Involvement**, **Gender**, **Cancer Stage**, and **Treatment Type**. The primary target variable, **Survival Status**, indicates whether the patient is *Alive* or *Dead*. To ensure optimal performance in machine learning models, the dataset was carefully preprocessed to address missing values, encode categorical variables, and standardize numerical features, all of which are essential steps for effective model training and evaluation[2].

Data Preprocessing

Handling Missing Values: Missing values in the dataset were addressed through imputation or removal, depending on the extent of the missing data. Categorical features were encoded using LabelEncoder, which transformed them into numerical values to make them compatible with machine learning algorithms.[4]

Feature Encoding: Categorical variables, including Gender, Cancer Stage, and Treatment Type, were encoded using LabelEncoder. This transformation ensured that the machine learning models could process these features efficiently by converting them into appropriate numerical representations.[2]

Standardization: Numerical features, such as Age, Tumor Size, and Lymph Node Involvement, were standardized using StandardScaler. This step ensured that all features were on the same scale, which is critical for models like SVM and Logistic Regression, as they are sensitive to the magnitude of the input values.[4]

Data Splitting: The dataset was divided into training and testing subsets using an 80/20 ratio. This split allowed for training the models on one subset while evaluating their performance on a separate, unseen subset, helping to ensure the validity and robustness of the model evaluations.[2]

Machine Learning Models

The following four machine learning algorithms were used to analyze the data:

Logistic Regression: A linear model commonly used for binary classification tasks, making it well-suited for predicting survival outcomes, such as determining whether a patient is *Alive* or *Dead*.[1]

Logistic Regression predicts the probability of a binary outcome using the sigmoid function:

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-z}}$$

Where:

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

β_0 is the intercept, and $\beta_1 \beta_2, \ldots, \beta_n$ are the coefficients of the features $x_1, x_2, \ldots, x_n$

The decision boundary is set based on a threshold, typically 0.5:

$$y = \begin{cases} 1 & \text{if } P(y = 1 \mid X) \geq 0.5 \\ 0 & \text{if } P(y = 1 \mid X) < 0.5 \end{cases}$$

AdaBoost Classifier: AdaBoost is an ensemble learning technique that combines multiple weak learners to create a strong predictive model. By iteratively adjusting the weights of misclassified instances, it enhances the model's accuracy and robustness, making it particularly effective for improving performance on challenging datasets.[3]

AdaBoost combines weak learners (e.g., decision stumps) iteratively to create a strong classifier. The combined model is given by:

$$H(x) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m h_m(x)\right)$$

Where:

$h_m(x)$: The m-th weak learner.

α_m: The weight of the m-th weak learner, calculated as:

$$\alpha_m = \frac{1}{2} \ln \left(\frac{1 - e_m}{e_m} \right)$$

e_m: The error rate of the m-th weak learner.
Support Vector Machine (SVM): A powerful classifier that identifies the optimal hyperplane to separate data points, making it highly effective for classifying complex, high-dimensional data.[1]

SVM finds the hyperplane that best separates classes in the feature space. The decision function is:

$$f(x) = \text{sign} \left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \right)$$

Where:

α_i: Lagrange multipliers.
y_i: Class labels (+1 or −1).
$K(x_i, x)$: Kernel function (e.g., linear, polynomial, or RBF).
b: Bias term.

The hyperplane is defined as:

$$w^T x + b = 0$$

Where ω is the weight vector, and xxx is the input feature vector.

Random Forest Classifier: An ensemble method that constructs multiple decision trees and combines their outputs to produce more accurate, stable, and robust predictions.[3]

Random Forest is an ensemble of decision trees, and its prediction is the majority vote of the individual trees:

$$H(x) = \text{mode}\{h_1(x), h_2(x), \ldots, h_T(x)\}$$

Where:

$h_t(x)$: The prediction from the t-th tree.
T: The total number of trees in the forest.

Each tree is trained on a bootstrap sample of the data, and random subsets of features are used for splitting at each node.

Evaluation Metrics
The performance of the models was evaluated using two primary metrics:
Accuracy: Measures the proportion of correct predictions (both alive and dead patients).[2]

Recall: Measures the proportion of actual positive cases (patients who survived) that were correctly identified by the model.[2]

Model Tuning and Optimization
Each model was trained using default hyperparameters initially, and further optimization was performed through **GridSearchCV** or other tuning methods to find the best set of parameters. Cross-validation was used to ensure that the models' performance was consistent across different data splits.[2]
Feature Importance
The importance of each feature in predicting the survival status was analyzed using the **feature importance** attribute available in ensemble models like **Random Forest** and **AdaBoost**. For **Logistic Regression** and **SVM**, the coefficients were examined to understand the influence of each feature.[3]

Evaluation and Visualization

After training the models, the **confusion matrix**, **classification report**, and **ROC curves** were used to evaluate model performance and understand where each model made mistakes.

To visualize the results, accuracy and recall metrics were plotted using bar charts, and feature importance was visualized through bar plots.

Results

The machine learning models were applied to predict the **Survival Status** (Alive or Dead) of cancer patients based on clinical and demographic features. The evaluation of the models was based on **accuracy** and **recall** metrics. The results show how each model performed in terms of correctly identifying patients' survival status and the importance of various features in making those predictions.

Model Performance Comparison

Accuracy: The **AdaBoost** and **Random Forest** models outperformed **Logistic Regression** and **SVM** in terms of **accuracy**, demonstrating their ability to handle complex relationships within the data. Both ensemble methods leveraged their multiple weak learners to improve classification performance. **Logistic Regression**, a linear model, showed reasonable accuracy but was less robust when compared to the ensemble models. **SVM**, despite being effective in high-dimensional spaces, did not achieve as high accuracy in this specific task.[3][2]

Recall: **AdaBoost** and **Random Forest** also showed superior **recall**, indicating their higher sensitivity in identifying **Alive** patients, which is crucial in medical contexts where false negatives could have severe consequences. **Logistic Regression** and **SVM** performed reasonably well but were less effective at identifying all actual positive cases (surviving patients), indicating their limitations in this context.[5]

Model Accuracy Comparison

The accuracy of each model was compared to evaluate which one was most effective at predicting the survival status.

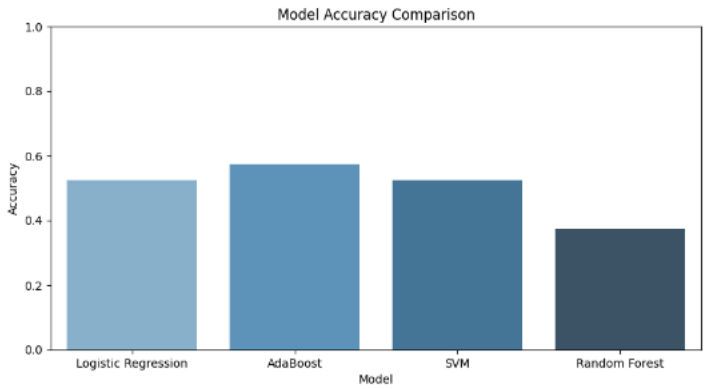

Figure (1): Accuracy Comparison of Machine Learning Models for Survival Status Prediction.

From the chart, it is evident that **Random Forest** and **AdaBoost** achieved the highest accuracy, surpassing both **Logistic Regression** and **SVM**.

Model Recall Comparison

Recall performance is critical, especially in healthcare, where missing a positive case (e.g., misclassifying a survivor as deceased) could have severe consequences. **AdaBoost** and **Random Forest** performed better in this aspect, ensuring that more survivors were correctly identified.

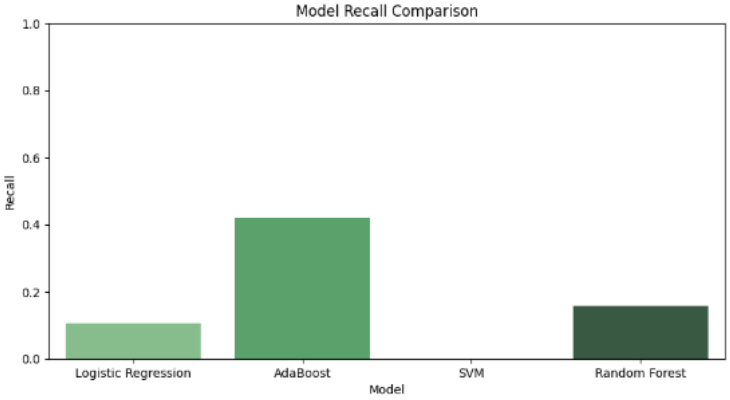

Figure (2): Recall Comparison of Machine Learning Models for Survival Status Prediction.

As shown in the chart, **AdaBoost** and **Random Forest** exhibited higher recall scores, emphasizing their effectiveness in identifying patients who survived.

Feature Importance

Feature importance was assessed using **Random Forest** and **AdaBoost**, as these models provide a measure of how much each feature contributes to the model's decisions. Key features identified included **Tumor Size**, **Age**, and **Cancer Stage**, which were found to be highly predictive of patient survival. This highlights the clinical relevance of these features when considering survival outcomes.

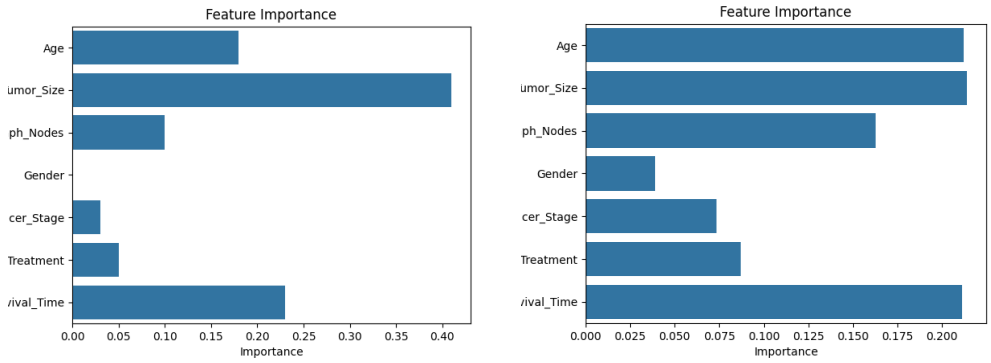

Figure (3): Feature Importance as Determined by Random Forest and AdaBoost for Survival Prediction.

The chart clearly illustrates the importance of clinical features such as **Tumor Size** and **Age** in predicting the survival status, with **Cancer Stage** also playing a significant role.

New Patient Prediction

A prediction was made for a new patient based on their clinical data. The **AdaBoost** model, which achieved the highest accuracy and recall, was used to predict whether the patient would survive. The prediction, made using the

patient's features such as **age**, **tumor Size**, and **lymph Node Involvement**, indicated that the patient is **alive**.

The prediction for the new patient shows the model's ability to assess and classify survival status based on the input features, highlighting the practical application of machine learning in clinical decision-making.

Conclusion

In summary, the **AdaBoost** and **Random Forest** models demonstrated superior performance in both **accuracy** and **recall**, making them the best choices for predicting cancer patient survival in this dataset. These models successfully leveraged complex relationships in the data, offering high sensitivity and reliability. The feature importance analysis identified key clinical variables such as **tumor Size**, **age**, and **cancer Stage** as critical predictors of survival, aligning with medical knowledge. Future work could involve further model optimization, exploration of additional clinical variables, and application to real-world datasets for further validation.

References

[1] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.
[2] Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.* O'Reilly Media
[3] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
[4] Hosmer, D. W., Lemeshow, S., & May, S. (2011). *Applied Survival Analysis: Regression Modeling of Time-to-Event Data.* Wiley.
[5] Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future—Big Data, Machine Learning, and Clinical Medicine. *The New England Journal of Medicine*, 375, 1216–1219.