Predictive Modeling of Patient Survival Outcomes in Immune Checkpoint Inhibitor Therapy

By. Young Beum (Ryan) Cho

Goal

(1) Derive adverse or beneficial survival outcomes of pertinent gene mutation statuses. Elucidate how they possibly contribute to immunotherapy response. (2) Build a machine-learning model that predicts patient survival outcomes. Compare performance across cancer types (3) Elucidate interactions between the predictors.

Data Preprocessing

The preprocessing function prepares the dataset for machine learning by identifying outliers, calculating gene mutations, converting categorical variables to numerical, and normalizing numerical columns. After preprocessing, the dataset is further refined by replacing the single 'Skin Cancer, Non-Melanoma' instance with 'Cancer of Unknown Primary', splitting the dataset while maintaining 'CANCER_TYPE' proportions, and creating a separate test set for the lone 'Skin Cancer, Non-Melanoma' case.

Logistic Regression

Logistic regression was used to determine the impact of gene mutation statuses on survival outcomes (Figure 1A), given its suitability for binary outcomes like survival or death. The model learns the relationship between features (e.g., gene mutations, TMB score) and the target variable (OS_STATUS). The 'train_and_evaluate' function encapsulates model training, prediction, and evaluation.

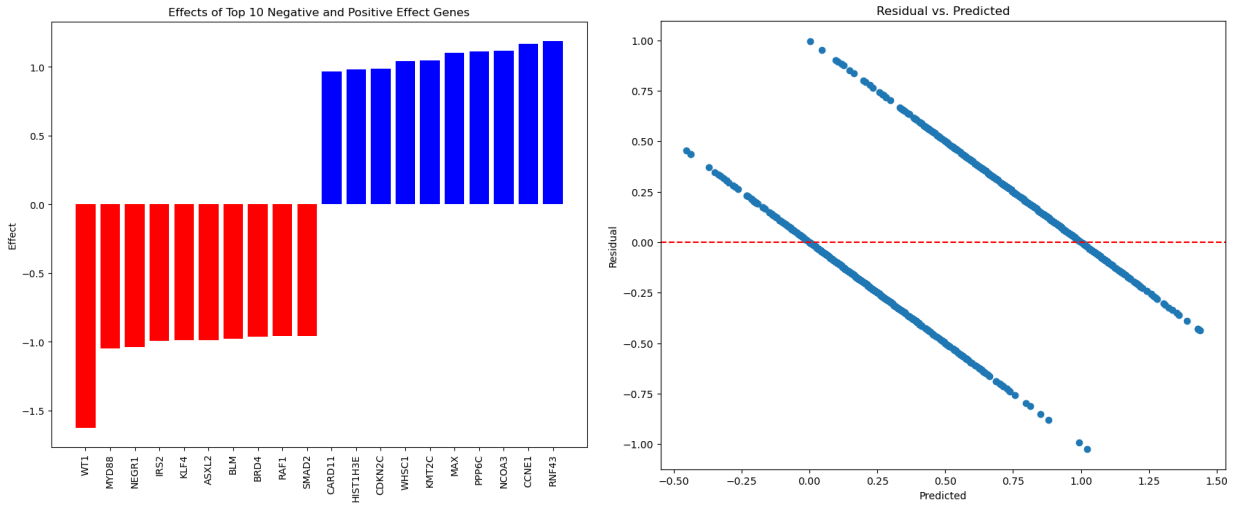


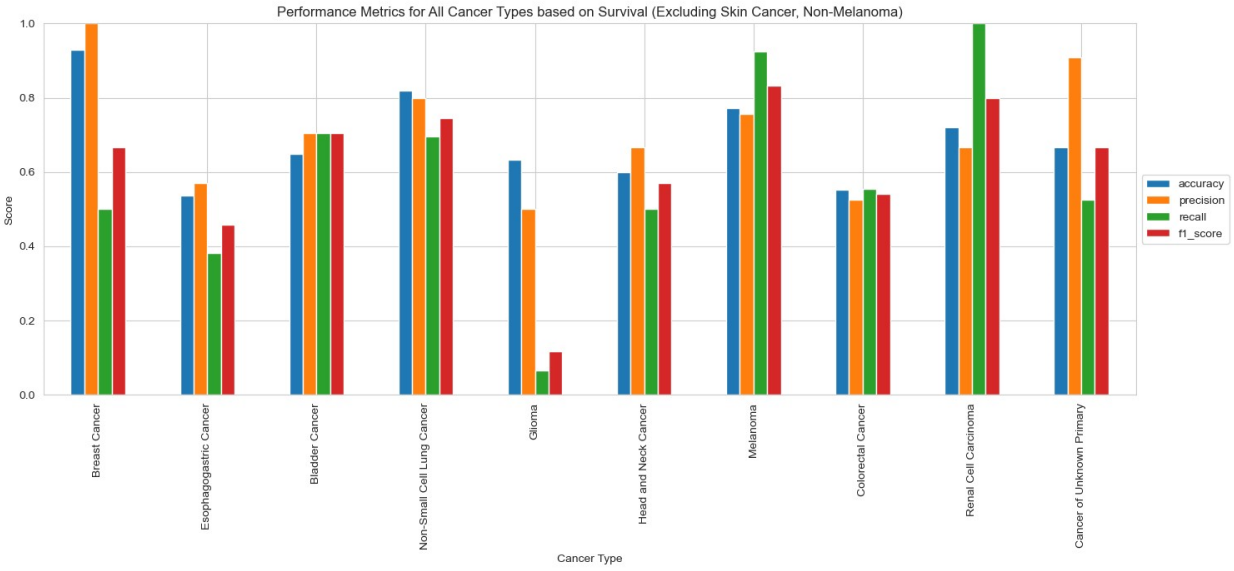
Figure 1A, B: Top 10 Negative and Positive Effect Genes (Left), Residual VS Predicted (Right)

To elucidate the effects of gene mutations on immunotherapy response, a quick literature fetching technique provided by (https://github.com/ybryan95/PubMed_scraper_GPT) was implemented to acquire relevant articles about genes with top impact using the query: *"(" + gene_name + "[Title/Abstract]) AND ((PD-1[Title/Abstract]) OR (PDL-1[Title/Abstract]) OR (CTLA4[Title/Abstract])) AND (Cancer[Title/Abstract])").* Additionally, the GPT model was implemented to use extracted article abstracts to elucidate gene mutation effects (Supplementary Data: textInsight_Abstract). However, as seen in the Residual VS Predicted plot, (Figure 1B) a systematic bias could be inferred. Thus, a non-linear predictive model was further considered for implementation.

Random Forest Classifier (RFC)

RFC was selected for its robustness and ability to handle complex, non-linear datasets like gene mutations. It reduces overfitting and improves accuracy by aggregating results from multiple decision trees. The 'train_and_evaluate' function streamlines model training, prediction, and evaluation, and identifies top contributing features for insights into factors influencing patient survival.
Performance metrics across all the cancer types can be found in Figure 2A (skin cancer excluded). Individual confusion matrices and model performance for all cancer types, and variables with high feature importance are provided in the supplementary data. However, due to the high-dimensional nature of the data, the max feature importance to be found was ‘OS_MONTHS’, with a value of 0.12. This indicates the model is likely to predict by a combination of many variables, rather than with one or several contributions from high-importance variables.

Figure 2: Model Performance across Cancer Types



Correlations between the Predictors

Relevant gene relationships were analyzed using a correlation matrix and visualized through a heatmap. The matrix was computed using pandas, selecting only gene pairs with a correlation above 0.5. This data was then transformed into a heatmap using seaborn, with gene pairs as indices and correlation as values (Supplementary Data). The heatmap's color scheme intuitively indicated the correlation's strength, aiding in understanding biological processes and treatment development.

Closure

In summary, while the models showed encouraging results (67% Median, 71% Mean Accuracy for RFC), they could be improved with a larger, balanced dataset across all cancer types. Synthetic data generation techniques like bootstrapping could enrich the dataset, but skin cancer should be excluded to avoid overfitting. Incorporating 'OS_MONTHS' in survival analysis could offer deeper insights into patient outcomes. Further feature engineering, such as indicating co-mutations in multiple genes, could also enhance model performance.