Alejandro Scaffa

CS5100

Apr 22, 2025

Final Project Report: PAN-Cancer ML Analysis

Introduction:

Cancer: Cancer is one of the deadliest groups of diseases in the world, with lung cancer, colorectal cancer, liver, and pancreatic cancer being the biggest killers. As per the World Cancer Research Fund, in 2020 there were 1.8 million lung cancer deaths, >900K colorectal cancer deaths, and >800K liver cancer deaths across the entire world (WHO, 2020), with >600K cancer deaths expected every year from the USA alone (WCRF, 2025). A subset of scientists studying how to treat cancer more efficiently and effectively are those in the field of transcriptomics, who take RNA-sequencing data from patients with and without cancer and try to find pertinent information that diagnose the cancer as well as give insights in how to treat it.

Central Dogma (DNA-RNA-Protein): As computer scientists, it is important to explain what RNA is in a simplified manner. So here is a primer: DNA is the recipe book with all the recipes that make you who you are. RNA is how often each page of the recipe book is being used. Proteins are the final recipe as food. DNA is used to make RNA, which in turn is used to make proteins - a process known as the central dogma of biology. Proteins are the most informative but the hardest and most expensive to study, especially in bulk. DNA, RNA, and protein are all useful information, and we read or "sequence" them in order to actually acquire the information from a sample. In general, biologists often think of RNA a bit more than DNA because RNA can offer information on how the DNA is being used (or not), and how that is having effect on phenotype/outcome of the disease/patient. An intuitive yet oversimplified example is that you are born with the DNA for the proteins that you need both as a kid and you as an adult. For example, both fetal (baby) and adult hemoglobin genes are encoded in DNA, but only the appropriate one is expressed at a given age - something only RNA data can reveal. Similarly, RNA expression profiles differ between healthy and cancerous cells, and even between different types of cancer.

Unmet Need and Motivation: There is an unmet need for researchers to analyze and classify tissue biopsies that have not been scored by a physician. The significance of a reproducible tool that can do this classification (that is trained on gold standard data) is that researchers like myself would have more information about the genetic data that they are working with, being able to create better research plans and biological models. Furthermore, this could be useful to train datasets with unlabeled data, but that you would like to be labeled to do supervised ML pipelines.

Training Dataset and Scope: This project aims to create a reproducible machine learning (ML) pipeline for the TCGA Pan-Cancer (PANCAN) dataset, a dataset that includes RNA-sequencing data acquired from 11,060 patient tumor samples, covering >20,000 gene transcription levels for each, across 33 different cancer types.

Project Objectives: Here I demonstrated the reproducible EDA, ML training, and evaluation of 1) a Binary algorithm that determines whether the RNA-sequencing sample is from a healthy sample or cancer, and 2) a Multi-classification algorithm which determines which kind of cancer out of the 33 possible options it was trained on. The final accuracy against the test sets was 99.3% for the binary classifier and 97.6% for the multiclass classifier.

Methods and Approach:

Data Acquisition: The genomic data was downloaded from the TCGA repository and the metadata was downloaded from the Xena Browser for the same PANCAN dataset. The dataset can be found in TCGA or here.

Data Pre-processing: Null values in the metadata were handled by removing entire rows with missing patient information, while nulls in the gene expression matrix were imputed as zeros (as recommended by domain experts, since undetected gene expression is equivalent to zero in this context). Duplicates were addressed, overall the file was very clean and had no more than 5% incomplete data. Sample type (cancer or not) and cancer type were binary and label encoded. The datasets were then merged, enabling the integration of metadata features such as age and gender into the ML pipeline alongside gene expression data. This is where the issue with class imbalance (90% tumor samples, vs 10% non tumor) was noticed, leading to a binary classifier that determines whether a sample is tumor or not, and a multiclass classifier that determines which kind of cancer that it is.

Data Analysis and Evaluation Strategy: The metadata revealed a strong class imbalance as mentioned above, which led to the design of two separate ML pipelines: a binary classifier to distinguish tumor vs. normal, and a multiclass classifier to predict one of 33 cancer types. Other metadata insights included a balanced gender distribution, a mean age of 65, and a relatively even spread across cancer types, making the dataset well-suited for multiclass learning.

To reduce dimensionality variance thresholding was applied to the gene expression data, and the best thresholds were selected based on cross-validated accuracy scores using a Random Forest classifier. The ML pipeline tested several off-the-shelf models, including logistic regression, random forest, naive Bayes, and XGBoost, using stratified 5-fold cross-validation. Scale standardization was applied using StandardScaler for logistic regression.

Model evaluation was based on multiple metrics: accuracy (overall performance), ROC AUC, precision, recall, and F1-score to study class-specific performance. Confusion matrices and ROC curves were generated to visualize prediction errors and classifier thresholds. All visualizations were created using matplotlib and seaborn, and the final trained models were saved using joblib for reproducibility and rapid re-use in the grading_document notebook.

Documents Provided: There are two jupyter notebooks provided with this final report, a file called eda_ML_cancer_genomics_final_scaffa.ipynb, and another called grading_document.ipynb. The first document took 4-5 days to run and contains the entire

pipeline from data acquisition to results. The second notebook re-tests all final models and evaluation metrics on the test data ensuring reproducibility and easy evaluation of performance in under 5 minutes.

Notebook Organization: The notebook is split into four parts. Part 1 covers EDA, including cleanup and analysis of the genomic and metadata. Part 2 includes ML prerequisites such as variance selection, off-the-shelf model comparisons, and data scaling. Parts 3 and 4 implement and evaluate the final binary and multiclass stacking classifiers.

Results and Analysis:

The reason for two pipelines and not one is that the dataset has a class imbalance of 90% tumor to 10% healthy, so the first pipeline will deal with this imbalance, but the second doesn't have to, leading to two stronger models when built separately. The metadata also showed a higher prevalence of breast cancer (BRCA) but still a good classification distribution across cancer types without much imbalance, a perfect gender balance, and a mean age of about 65 years old (Figure 1).

To address the high dimensionality of the gene expression data, variance thresholding was performed. As shown in Figure 2A, this step helped identify the most predictive features, with optimal thresholds of 0.049 for binary classification and 0.229 for multiclass classification, based on cross-validated performance. Following this, off-the-shelf models were trained and evaluated (Figure 2B), with logistic regression and XGBoost showing the best performance in both tasks.

In the binary classification task, logistic regression achieved 97.4% accuracy, XGBoost reached 99.2%, and the stacking model combining both peaked at 99.3% accuracy (Figure 3). Beyond accuracy, the final stacking model also achieved excellent scores in precision (0.995), recall (0.998), and F1-score (0.996). Lastly, the ROC AUC was of 0.991, which is particularly great because a class imbalance can lead to a low ROC AUC and the class prediction caused by biased towards the majority class (Figure 5).

For the multiclass classification task, logistic regression and XGBoost again performed strongest, with stacking slightly outperforming them at 97.6% accuracy (Figure 4). The confusion matrix (Figure 6) revealed that most cancer types were predicted with very high accuracy, although some rare classes were more frequently misclassified. To check on these, another confusion matrix for the 10 worst-performing cancer types was plotted separately (Figure 7A), showing where the model struggled. These are under-represented classes and the false positives mostly come from misclassifying particularly similar cancer types. Despite these challenges, the overall ROC AUC remained extremely high (0.9997) with about 97.5% values for precision, recall, and F1-score (Figure 7C).

Conclusions and Reflections:

Overall, this project was successful. After doing the EDA, exploring logistic regression, random forest, naive bayes, and XGBoost models, I settled on stacking models using the hyperparameter tuned version of the logistic regression and the XGBoost models for both the binary classifier and the multiclass classifier. The final stacking models achieved 99.3% accuracy on the binary classification task and 97.6% accuracy on the multiclass task, both were evaluated on held-out test sets. Meaning the goal of creating an ML pipeline that can accurately classify whether RNA-sequencing data is cancer (binary) and which kind of cancer out of 33 (multiclass) was achieved. There are some future directions for this work, and that is expansion of its use. The TCGA PANCAN dataset is of high quality and quite complete. Oftentimes you don't get all 20,000 genes sequenced reliably. I did variance thresholding to lower the number of genes needed for the model to train on, but I believe even less complete data is out there. So, I foresee a future direction including training a model with a much higher dimensionality reduction through variance thresholding and PCA analysis that requires less genes to make a prediction. Finally, another important step would be to grow this dataset, to merge this dataset with others to lower issues of class imbalance that were seen in the multiclass classifier. Particularly datasets for CHOL (bile duct cancer), RHEA (rectum cancer) and UCS (uterine cancer).

Bibliography:

"Cancer." *World Health Organization*, www.who.int/news-room/fact-sheets/detail/cancer. 2020.

"Global cancer data by country." *World Cancer Research Fund*, www.wcrf.org/preventing-cancer/cancer-statistics/global-cancer-data-by-country/. 2025.

Image Appendix:

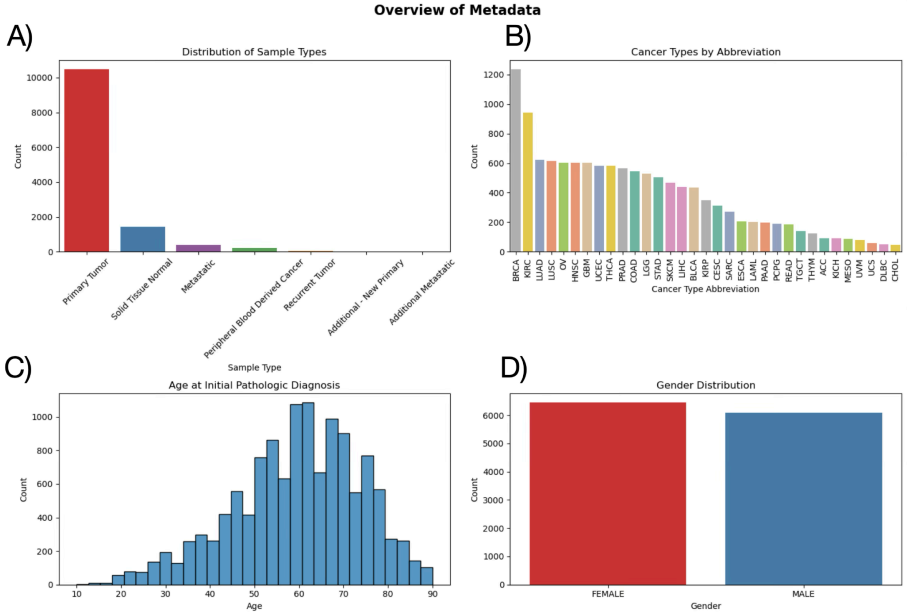

Figure 1: A) Distribution of sample types (solid tissue normal, and multiple types of cancer), B) Counts of cancer types by abbreviation throughout the dataset, C) Age at original diagnosis, and D) Gender Distribution.

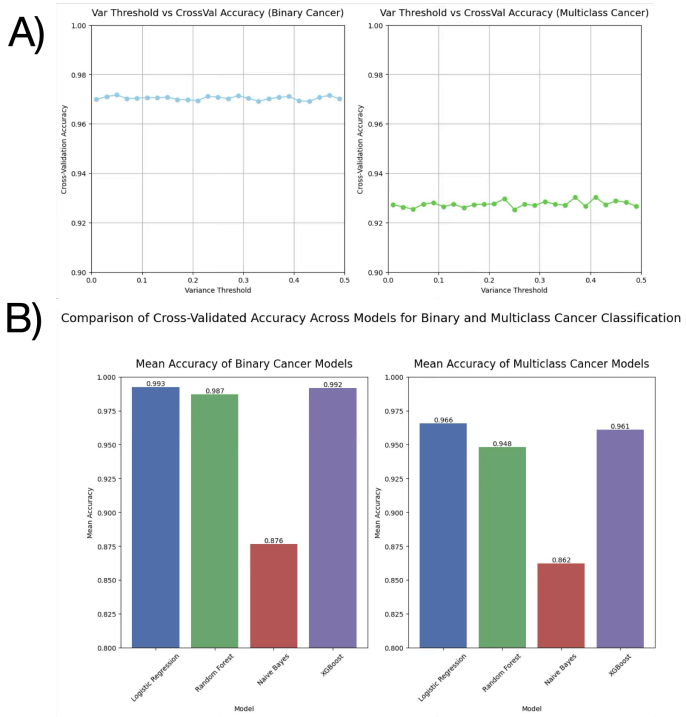

Figure 2: 2A) Variance Thresholding for Binary and Multiclass models graphs of cross-validation score vs variance threshold. Best variance 0.049 for binary and 0.229 for multiclass 2B) Mean accuracy results for binary and multiclass off-the-shelf models including logistic regression, random forest, naive bayes, and XGBoost. Logistic Regression and XGBoost were two best performers and selected for hyperparameter tuning.

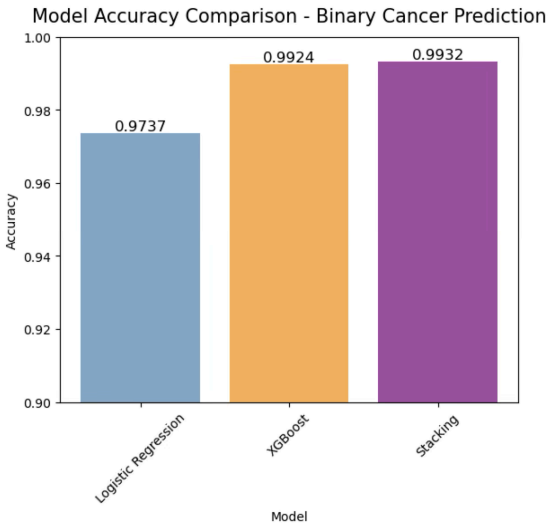

Figure 3: Model accuracy comparison for the binary classifier including hyperparameter optimized logistic regression (0.9737), XGBoost (0.9924), and stacking model of both (0.9932). Stacking model selected for final evaluation.

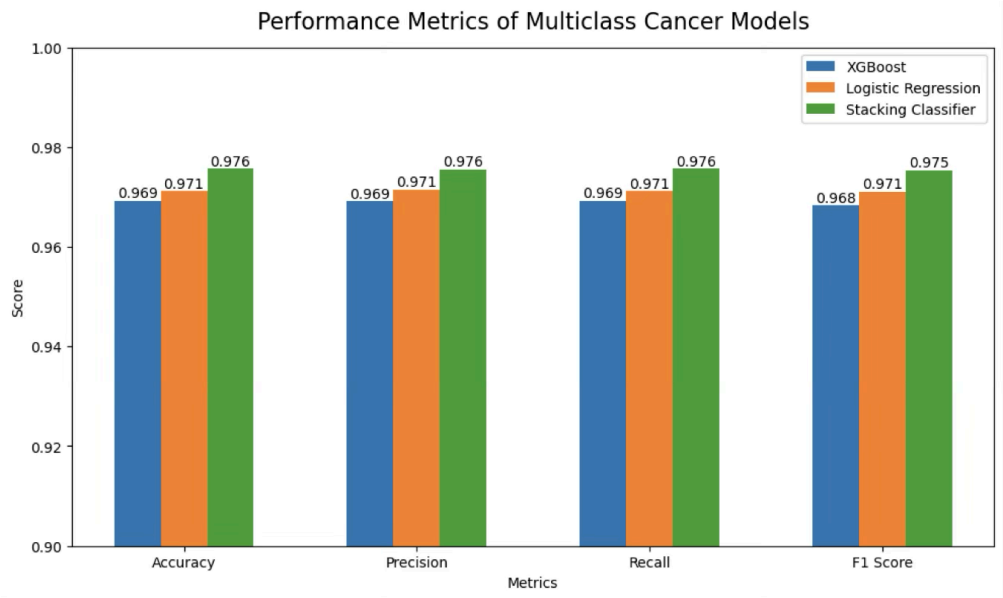

Figure 4: Model accuracy, precision, recall, and F1 score comparison for the multiclass classifier including hyperparameter optimized logistic regression accuracy of (0.969), XGBoost (0.971), and stacking model of both (0.976). Stacking model selected for final evaluation.

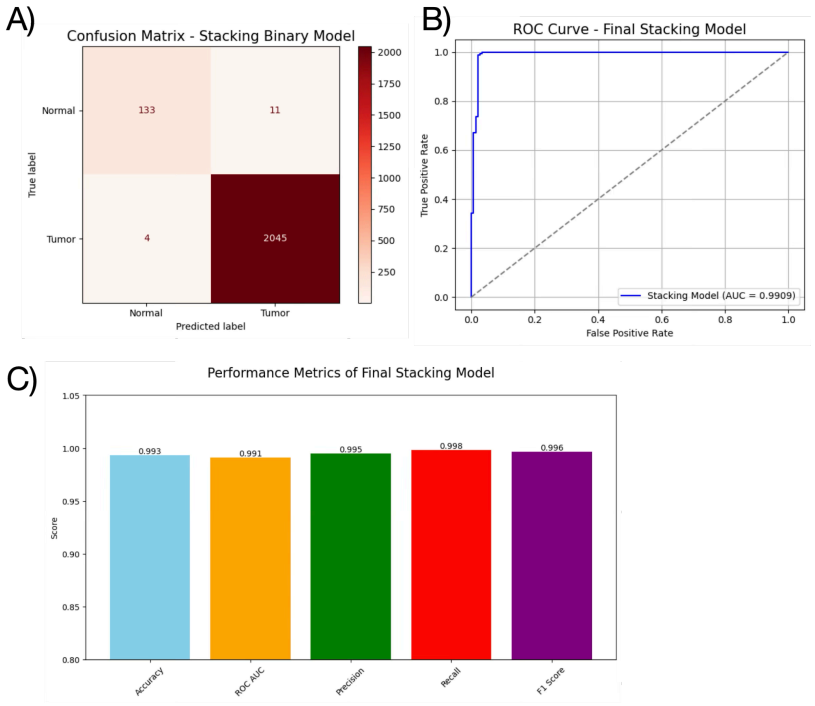

Figure 5: 2A) Confusion matrix for stacking (logistic regression and XGBoost) binary cancer prediction model on a test set of 20% total data. 2B) ROC Curve for Binary Classification Stacked Model. 2C) Accuracy (0.993), ROC AUC (0.991), Precision (0.995), Recall (0.998), and F1-Score (0.996) for Binary Classification Stacked Model

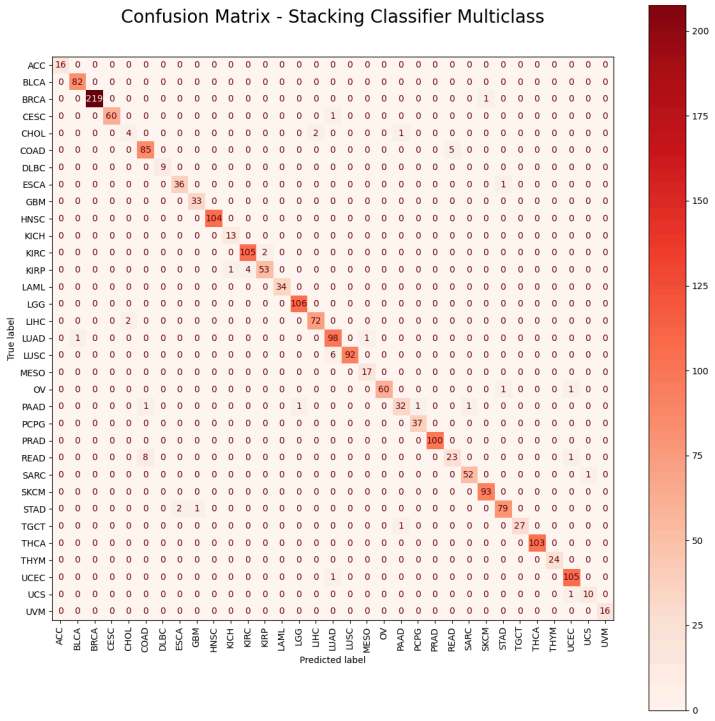

Figure 6: Confusion matrix for stacking logistic regression and XGBoost multiclass cancer prediction model on a test set of 20% total data.

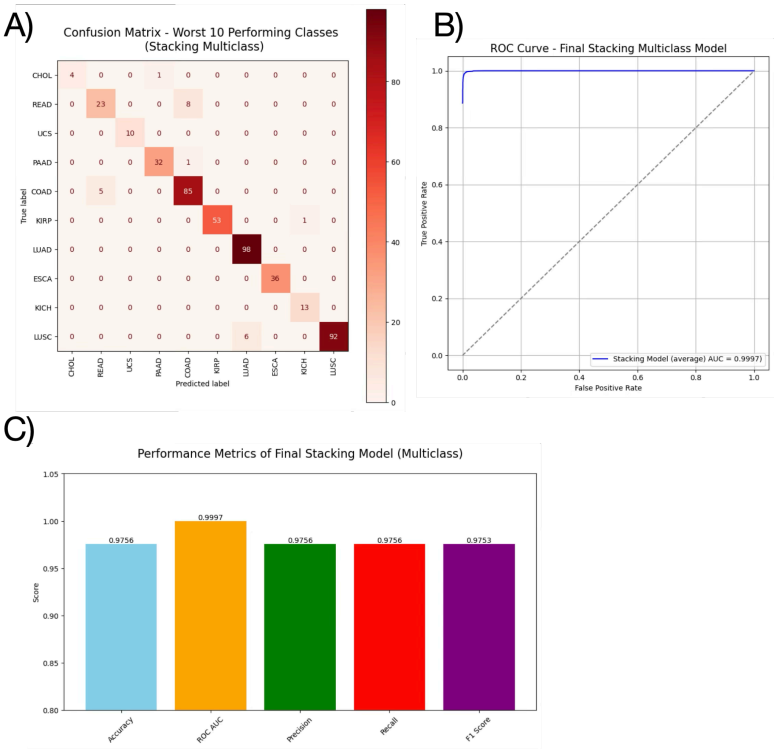

Figure 7: 4A: Confusion matrix for top 10 worst performers cancer types of the stacking classifier for the multiclass cancer prediction model on a test set of 20% total data. 4B: ROC Curve for Binary Classification Stacked Model. 4C: Accuracy (0.076), ROC AUC (0.9997), Precision (0.9756), Recall (0.9756), and F1-Score (0.9753) for Multiclass Classification Stacked Model.