Predictive Modeling for Mortality in a Pan-Cancer Cohort: MSK-MET Analysis.

Gagandeep Singh[#1], Akshar Patel[#2], Ankur Mangroliya[#3], Deon Lobo[#4]

[#]*Computer Science Department, University of Windsor, Canada*

[1]gagande1@uwindsor.ca

[2]patel2x8@uwindsor.ca

[3]mangrol3@uwindsor.ca

[4]lobod@uwindsor.ca

***Abstract*— This project leverages machine learning techniques to explore factors influencing overall survival in a diverse cohort of metastatic cancer patients. The analysis involves preprocessing clinical data from the MSK-MET dataset, addressing missing values, and conducting exploratory data analysis to understand metastases types' distribution and relationships with demographic factors. Principal Component Analysis (PCA) is applied for dimensionality reduction, capturing essential information and reducing dataset complexity. Subsequently, the study employs feature selection techniques, such as SelectKBest and Mutual Information, to optimize the dataset for predictive modeling. Logistic Regression, Support Vector Machine (SVM), and Random Forest Classifier models are then trained and evaluated, demonstrating their effectiveness in predicting overall survival status. This comprehensive approach contributes to the development of a robust predictive model for mortality in a pan-cancer cohort, offering valuable insights for clinical prognosis in metastatic cancer patients.**

***Keywords*— Predictive Modeling, Mortality, Pan-Cancer Cohort, Machine Learning, Feature Selection, Principal Component Analysis (PCA), MSK-MET Dataset, Metastatic Cancer, Clinical Prognosis, Overall Survival.**

I. Introduction

This document presents a comprehensive analysis of clinical data related to metastatic cancer patients. The dataset is preprocessed to handle missing values, remove irrelevant columns, and one-hot encode categorical variables. Exploratory data analysis is conducted to visualize the distribution of distant metastases and examine relationships between overall survival status and demographic factors such as sex, race category, and cancer type.

Following data exploration, principal Component Analysis (PCA) is employed for dimensionality reduction to capture essential information and reduce the complexity of the dataset in this analysis. Two feature selection techniques are employed: SelectKBest and Mutual Information. The reduced feature sets are used to train three distinct machine learning models—Logistic Regression, Support Vector Machine (SVM), and Random Forest Classifier. The models are evaluated based on their accuracy in predicting overall survival status.

This study showcases the application of machine learning techniques in predicting overall survival status for metastatic cancer patients, contributing valuable insights into the interplay between clinical features and patient outcomes. The methodologies presented can aid in the development of more efficient and interpretable models for cancer prognosis.

II. Description Of The Problem

The main problem addressed in this project revolves around predicting overall survival in a pan-cancer cohort, specifically within the context of metastatic cancer patients. In the realm of oncology, understanding the factors influencing mortality is a pivotal concern with substantial implications for treatment strategies and patient care. The distinctiveness of our solution lies in its comprehensive approach to handling an extensive dataset comprising 25,000 patients, a challenge attributed to the data's complexity and scale, rendering traditional analysis methods cumbersome. Our solution encompasses meticulous data preprocessing, adeptly managing missing values, and a thorough exploratory data analysis. Advanced machine learning techniques for

feature selection and dimensionality reduction are integral components, revealing intricate patterns and relationships within the dataset. This collaborative and multi-faceted approach provides a nuanced understanding of the factors contributing to overall survival. By integrating expertise in data science, machine learning, and domain knowledge, our solution presents a unique and comprehensive methodology to address a significant issue in cancer research, offering insights that can substantially impact clinical decision-making and patient outcomes.

III. Role Of Group Participants

In this collaborative project, each participant played a crucial role in ensuring the success of the endeavor. Ankur took charge of handling missing values and refining the dataset to ensure its quality and reliability, leveraging expertise in data preprocessing and cleaning. Akshar, specializing in exploratory data analysis, conducted an in-depth examination of the dataset, exploring the distribution of metastases types, and establishing relationships with demographic factors. Gagandeep, with a focus on machine learning and feature selection, implemented techniques such as SelectKBest and Mutual Information to optimize the dataset for predictive modeling. Additionally, Gagandeep and Ankur collaborated with Deon, in implementing dimensionality reduction techniques like Principal Component Analysis (PCA), enhancing the efficiency of the dataset for machine learning models. Deon, Ankur, Akshar and Gagandeep actively participated in training and evaluating the models, creating graphs and ensuring their effectiveness in predicting overall survival status. Through this equitable distribution of tasks, the team maximized the diverse skills of each participant, fostering a collaborative and efficient work environment.

IV. Solution And Discussion

In our discussion, we'll delve into the details of our solution, encompassing meticulous data preprocessing for a large dataset and advanced techniques such as exploratory data analysis, feature selection, and dimensionality reduction.

A. Libraries

In the pursuit of gaining profound insights into clinical data related to metastatic cancer patients, a diverse set of libraries is employed to facilitate data manipulation, visualization, dimensionality reduction, and machine learning model implementation. This ensemble of libraries serves as a robust toolkit to unlock patterns, relationships, and predictive capabilities within a complex dataset.

1) Pandas: Pandas, a fundamental library in the Python ecosystem, is employed for data manipulation and analysis. Its powerful DataFrame structure allows for efficient handling of tabular data, facilitating tasks such as loading the clinical dataset, handling missing values, and performing preliminary data exploration.

2) Seaborn and Matplotlib: Seaborn and Matplotlib are indispensable for data visualization. Seaborn, built on top of Matplotlib, offers a high-level interface for aesthetically pleasing statistical graphics. Matplotlib, on the other hand, provides a comprehensive set of plotting functionalities. Together, they enable the creation of informative and visually appealing plots, aiding in the exploration of the dataset's distribution, relationships, and trends.

3) Scikit-learn (sklearn): Scikit-learn is a versatile machine learning library that plays a pivotal role in this analysis. It provides a wide array of tools for data preprocessing, model selection, and evaluation. The "PCA" module from sklearn is utilized for Principal Component Analysis, a technique employed for dimensionality reduction. Additionally, "train_test_split" assists in splitting the dataset into training and testing sets, while classifiers like "RandomForestClassifier," "LogisticRegression," and "SVC" are leveraged for predictive modeling.

4) RandomForestClassifier, LogisticRegression, and SVC: These classifiers are key components for training machine learning models. RandomForestClassifier is an ensemble learning method that constructs multiple decision trees and merges them to improve predictive accuracy and control overfitting. LogisticRegression is a linear model used for binary classification, suitable for predicting outcomes like survival status. Support Vector Classifier (SVC) is effective for both binary and multiclass classification, offering flexibility in handling diverse clinical outcomes.

5) Accuracy_score: The accuracy_score metric from sklearn is employed to assess the performance of the machine learning models. It quantifies the accuracy of predictions by comparing the predicted labels with the true labels in the testing set.

B. Dataset

The dataset, acquired from the cBioPortal study titled "MSK-MET (Memorial Sloan Kettering - Metastatic Events and Tropisms)," serves as a comprehensive pan-cancer cohort encompassing genomic and clinical outcome data from a staggering 25,000 patients. Analyzed with the objective of unraveling the genomic mechanisms driving metastatic progression, the study reveals intriguing associations between genomic alterations and patterns of metastatic dissemination across 50 diverse tumor types.

The significance of this dataset extends beyond the identification of associations. The study sheds light on the specific somatic alterations linked to increased metastatic burden and discerns patterns associated with distinct routes of metastatic spread. The data from MSK-MET thus emerges as a valuable resource, offering a unique platform for delving into the biological underpinnings of metastatic spread. Notably, the study underscores the pivotal role of chromosomal instability in cancer progression, providing nuanced insights that hold promise for advancing our understanding of metastatic mechanisms and influencing therapeutic strategies in the realm of oncology.

C. Remove columns that are not helpful in prediction

To streamline the dataset for effective analysis and prediction, redundant and non-informative columns were identified and removed from the original dataset. The columns "Study ID," "Patient ID," "Sample ID," "Number of Samples Per Patient," "Cancer Type Detailed," and "Subtype Abbreviation" were deemed unnecessary for the analytical objectives. These columns primarily consisted of identifiers and redundant information that does not contribute significantly to predictive modeling. By eliminating these non-informative columns, the resulting dataset, represented by the variable og_df_removed, is optimized for focused analysis, enhancing computational efficiency and ensuring a more meaningful exploration of the relationships between relevant features and the target variable during the subsequent stages of the analysis.

1) StudyID, PatientID, SampleID, Cancer Study: These are just Ids and are of no use for analysis and prediction.

2) Number of Samples Per Patient: Its value is 1 mostly and looks like it will not give any advantage during prediction.

3) Cancer Type Detailed: There is already one field of Cancer Type so this is redundant.

4) Subtype Abbreviation: There already exists a SubType column so this is redundant.

D. Statistical overview of data

	Age at Death	Age at First Mets Dx	Age at Last Contact	Age at Sequencing	Age at Surgical Procedure	FGA	Fraction Genome Altered	Met Count	Met Site Count	MSI Score	Mutation Count	Overall Survival (Months)	Sample coverage	TMB (nonsynonymous)	Tumor Purity
count	10023.000000	19638.000000	15646.000000	25517.000000	24953.000000	25775.000000	25775.000000	25775.000000	25775.000000	25773.000000	24753.000000	25659.000000	25775.000000	25775.000000	25368.000000
mean	64.897442	61.014257	63.576261	62.302248	61.513569	0.194624	0.194728	3.857187	3.011988	1.285071	9.208217	22.112903	660.395267	7.954993	39.735375
std	12.564958	12.859221	12.942135	12.862323	12.943476	0.191566	0.191612	3.736549	2.579182	4.635967	22.262670	17.695178	212.485348	19.447016	20.680170
min	19.020000	12.570000	18.410000	18.290000	17.160000	0.000000	0.000000	0.000000	0.000000	-1.000000	1.000000	0.000000	100.000000	0.000000	0.000000
25%	57.160000	53.072500	55.552500	54.390000	53.560000	0.030000	0.029900	1.000000	1.000000	0.000000	3.000000	8.440000	526.000000	2.218311	20.000000
50%	66.210000	62.280000	65.075000	63.700000	62.880000	0.143000	0.142800	3.000000	2.000000	0.230000	5.000000	17.450000	653.000000	4.323490	40.000000
75%	73.865000	70.130000	72.887500	71.460000	70.660000	0.305000	0.305400	6.000000	5.000000	0.860000	8.000000	33.120000	790.000000	6.917585	50.000000
max	90.000000	90.000000	90.000000	90.000000	90.000000	1.000000	1.000000	31.000000	15.000000	53.440000	758.000000	80.890000	2778.000000	654.576458	100.000000

Fig. 1 Getting Statistical overview of data

The dataset consists of 25,775 entries, providing insights into key features such as patient age at various milestones, genomic metrics like Fraction Genome Altered (FGA) and Mutation Count, metastatic counts, and overall survival metrics. The mean age at different events ranges from 61 to 64 years, with standard deviations indicating moderate variability. Notably, the mean FGA is approximately 19.5%, suggesting a considerable genomic alteration across the cohort. Metastatic counts exhibit a mean of 3.86, with a maximum of 31, indicating varying degrees of metastatic involvement. The overall survival ranges from 0 to 80.89 months, reflecting the diverse nature of clinical outcomes in the cohort. These statistical summaries lay the foundation for further in-depth exploration and modeling of the clinical data.

E. Checking for Null and Unique Values

The examination of unique values in each column of the preprocessed dataset, og_df_removed, provides insights into the diversity of categorical variables and the cardinality of continuous ones. Notably, columns like "Age at Death" and "Metastatic Site" showcase a range of unique values, indicating the variability in patient outcomes and metastatic involvement. Additionally, the assessment of missing values reveals that some columns, such as "Age at Death" and "Metastatic Site," exhibit a considerable percentage of missing data, reaching up to 61%. These missing values can impact the accuracy of predictive modeling, and addressing them appropriately is crucial. The analysis also highlights features with no missing values, such as "Cancer Type" and "Met Count," which provide complete information. This comprehensive examination of unique and missing values sets the stage for informed decision-making regarding imputation strategies and the selection of features for subsequent analysis and prediction tasks.

	null_values	null_value_percentage
Age at Death	15752	61.0
Metastatic Site	15632	61.0
Age at Last Contact	10129	39.0
Age at First Mets Dx	6137	24.0
Mutation Count	1022	4.0
Age at Surgical Procedure	822	3.0
Tumor Purity	407	2.0
Age at Sequencing	258	1.0
Overall Survival (Months)	116	0.0
Race Category	77	0.0
Sex	4	0.0
Primary Tumor Site	2	0.0
MSI Score	2	0.0
MSI Type	1	0.0
Oncotree Code	0	0.0
Metastatic patient	0	0.0
Cancer Type	0	0.0
Met Count	0	0.0
Met Site Count	0	0.0
Distant Mets: Biliary tract	0	0.0

Fig. 2 Null value and null percentage for each column

F. Visualizing and Handling Null Values

In the data preprocessing phase, a threshold of 30% was set to identify columns with a significant proportion of missing values, and those exceeding this threshold were dropped from the dataset, as imputing such a large amount of missing data may introduce bias. For numerical features like age, where missing values were observed, the mean value was calculated and used for imputation to maintain data integrity. The SimpleImputer module from scikit-learn facilitated this process. On the other hand, for categorical data, a new category labeled "unknown" was introduced to handle missing values in categorical columns, ensuring that the absence of categorical information is explicitly represented in the dataset. This approach allows for a balanced treatment of missing data in both numerical and categorical contexts, preserving the overall quality and representativeness of the dataset for subsequent analysis and modeling.

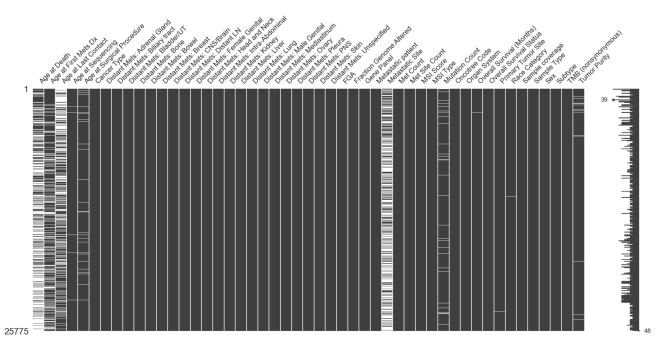


Fig. 3 Distribution of missing values across different columns

The matrix plot in Fig. 3 provides a quick and visual overview of the distribution of missing values across different columns. Each row corresponds to an entry in the dataset, and columns represent features. Non-empty cells are filled with color, while missing values are depicted as white spaces. This visualization aids in identifying patterns and clusters of missing data, offering insights into potential correlations between the absence of values in different columns. The accompanying adjustment of the figure size enhances clarity for a more detailed examination of missing data patterns.

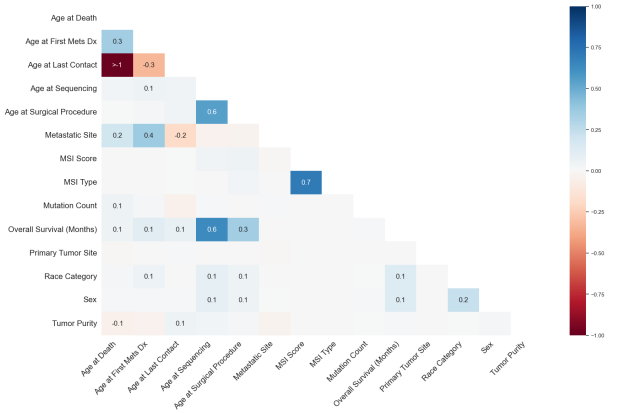


Fig. 4 Color-coded representation

The Fig. 4 heatmap provides a color-coded representation where darker squares indicate a higher correlation between the absence of values in pairs of columns. The inclusion of labels allows for a clear identification of specific columns involved in the correlations. This visualization is particularly useful for understanding potential relationships between missing values across different features, helping to inform imputation strategies and highlighting patterns in the missing data structure of the dataset.

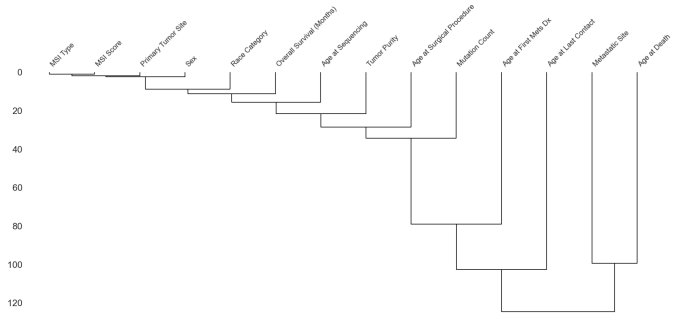


Fig. 5 Missing values in the preprocessed dataset

The Fig. 5 dendrogram illustrates hierarchical clustering of columns based on their patterns of missing data. Columns with similar patterns of missing values

are grouped together, and the vertical lines in the dendrogram represent clusters of features that share commonalities in the distribution of missing values. This visualization aids in understanding relationships between columns with missing data, potentially revealing underlying patterns or dependencies that can inform targeted strategies for imputation or further data analysis.

The missing data in the preprocessed dataset is addressed. Initially, it calculates the percentage of missing values for each column and sets a threshold of 30%. Columns exceeding this threshold are identified, resulting in the creation of the columns_drop index, which includes 'Age at Death', 'Age at Last Contact', and 'Metastatic Site'. Subsequently, these columns are dropped from the dataset, leading to the creation of a new DataFrame called df_selected_features.

To handle missing values in numerical columns, the code utilizes the SimpleImputer from scikit-learn with the strategy set to 'mean'. The 'Age at Sequencing', 'Age at Surgical Procedure', 'Age at First Mets Dx' and 'Age at Last Contact' columns are imputed with the mean values.

For categorical columns such as 'Sex' and 'Race Category', missing values are filled with the string 'Unknown' to designate the absence of information. Additionally, rows with missing values in crucial columns ('MSI Type', 'MSI Score', 'Primary Tumor Site', 'Overall Survival (Months)', 'Tumor Purity', and 'Mutation Count') are removed from the dataset, ensuring a more complete and reliable dataset for subsequent analysis. This comprehensive approach to handling missing data contributes to the robustness of the dataset for downstream machine learning tasks.

	null_values	null_value_percentage
Age at First Mets Dx	0	0.0
Oncotree Code	0	0.0
FGA	0	0.0
Fraction Genome Altered	0	0.0
Gene Panel	0	0.0
Metastatic patient	0	0.0
Met Count	0	0.0
Met Site Count	0	0.0
MSI Score	0	0.0
MSI Type	0	0.0
Mutation Count	0	0.0
Organ System	0	0.0
Age at Sequencing	0	0.0
Overall Survival (Months)	0	0.0
Overall Survival Status	0	0.0
Primary Tumor Site	0	0.0
Race Category	0	0.0
Sample coverage	0	0.0
Sample Type	0	0.0
Sex	0	0.0

Fig. 6 Missing values in dataset

This Fig. 6 presents the missing values in the df_selected_features DataFrame. It first computes the total number of missing values for each column, sorting them in descending order. The percentage of missing values for each column is then calculated and rounded to two decimal places. The results are combined into a DataFrame named null_value, where each row corresponds to a column, displaying both the total number and percentage of missing values. The presented information provides a clear overview of the extent of missing data in the selected features dataset, aiding further analysis and decision-making on handling missing values if needed.

G. Handling Outliers in data

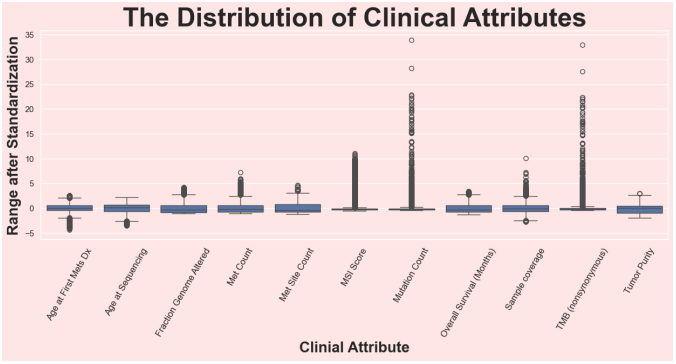


Fig. 7 Standardize numerical clinical attributes

This Fig. 7 standardizes numerical clinical attributes in the df_selected_features DataFrame using the Standard Scaler from scikit-learn. It selects specific numerical columns, scales them to have a mean of 0 and a standard deviation of 1, and then visualizes the distribution of these standardized attributes through a boxplot. The resulting plot provides insights into the range and distribution of each clinical attribute after standardization, aiding in the identification of potential outliers and the overall understanding of the dataset's numerical characteristics. The standardized values are plotted for each attribute, with the boxplots facilitating the comparison of their distributions. The visualization is enhanced with custom styling for clarity and emphasis. For example ‘Mutation count’ has a lot of outliers.

In this project identify and remove the outliers from the specified numerical columns (num_cols) in the df_selected_features DataFrame. For each column, it calculates the first quartile (q1), third quartile (q3), interquartile range (IQR), and sets lower and upper limits to define the range within which data is considered normal. Data points outside this range are

considered outliers and are subsequently removed from the DataFrame. The final DataFrame, df_removed_outliers, contains the data with outliers removed from the specified numerical columns.

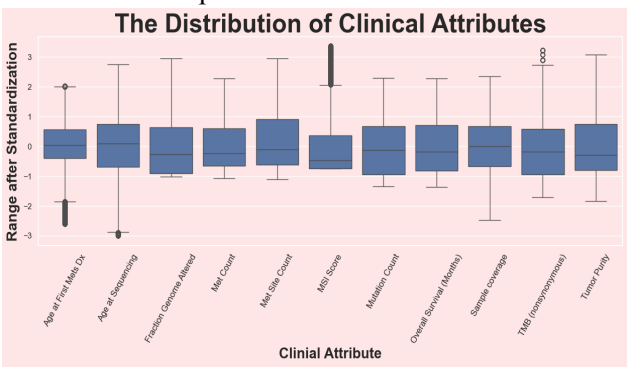


Fig. 8 Standardize numerical clinical attributes

The Fig. 8 box plot displays the central tendency, spread, and potential outliers in each standardized attribute, aiding in the assessment of data distribution and identifying patterns or variations in the clinical attributes.

H. Exploratory Analysis of Data

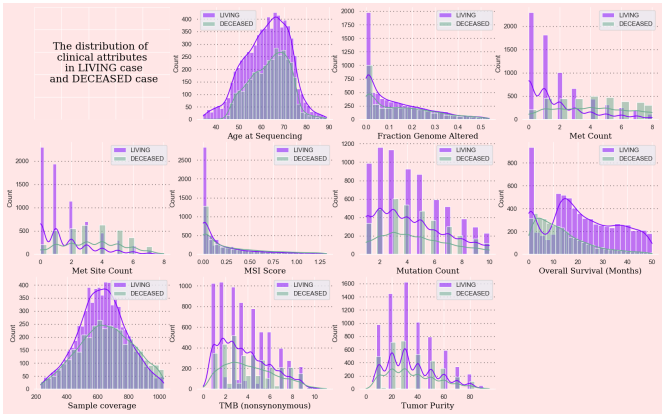


Fig. 9 Exploratory Analysis of Data

The Fig. 9 generates a grid of histograms to visualize the distribution of various numerical clinical attributes in both living and deceased cases. The histograms are plotted for each specified attribute, comparing the distribution of values between cases labeled as "LIVING" and "DECEASED" in the "Overall Survival Status" column. The color-coded histograms provide a visual representation of how the distribution of each clinical attribute varies between living and deceased cases. The resulting grid allows for a quick comparison of attribute distributions, aiding in identifying potential patterns or differences that may be relevant to overall survival outcomes. Additionally, the code includes annotations and formatting to enhance the clarity and interpretability of the visualizations.

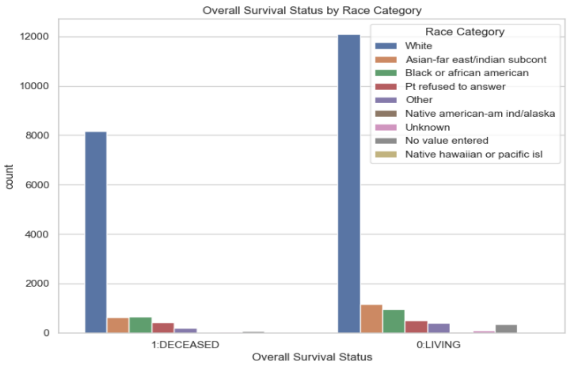


Fig. 10 Overall survival status by race category

Figure 10 illustrates a prominent demographic trend within the dataset, revealing a predominant representation of individuals belonging to the White Race, with Asians and other racial categories following suit. This observation underscores the importance of acknowledging and understanding the racial distribution in the data, as it can significantly impact the generalizability and interpretation of machine learning models trained on this dataset. Recognizing such demographic patterns is crucial for ensuring equitable and unbiased healthcare insights derived from the analysis of metastatic cancer patient data.

After experimenting with outlier removal, we observed a decrease in model accuracy, prompting the decision to exclude this preprocessing step in our final model training. Outliers, while potentially influencing certain statistical measures, can also contain valuable information about extreme cases or unique patient profiles in a medical dataset. By retaining these outliers, we aim to ensure a more comprehensive representation of the diverse clinical scenarios within metastatic cancer patients, promoting model generalization and robustness across the entire spectrum of cases.

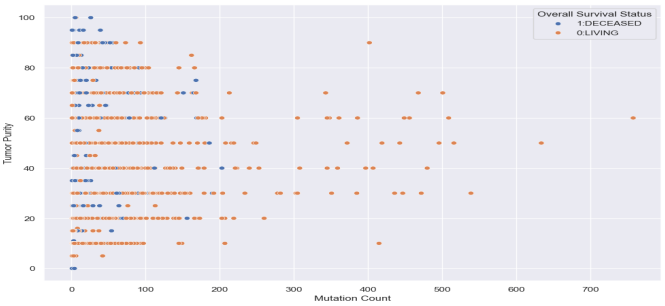


Fig. 11 Exploratory Analysis of Data

Fig. 11 generates a scatter plot using Seaborn, where the x-axis represents the "Mutation Count," the y-axis represents "Tumor Purity," and the data points are colored based on the "Overall Survival Status." Each point in the scatter plot corresponds to an entry in the dataset, and the hue (color) indicates whether the individual is categorized as "LIVING" or "DECEASED" in terms of overall survival status. The scatter plot allows for a visual examination of the relationship between mutation count, tumor purity, and overall survival status, helping identify potential patterns or trends in the data.

The conclusion drawn from the above graph suggests that there is a correlation between higher mutation counts and increased chances of survival. Additionally, it appears that cases with lower tumor purity, specifically below 70%, may be associated with a higher likelihood of survival. This observation underscores the potential significance of mutation count and tumor purity as factors influencing overall survival outcomes in the analyzed dataset. Further analysis and investigation into these patterns could provide valuable insights for understanding the dynamics of metastatic cancer progression.

I. Relationship between different cancer types and tumor purity

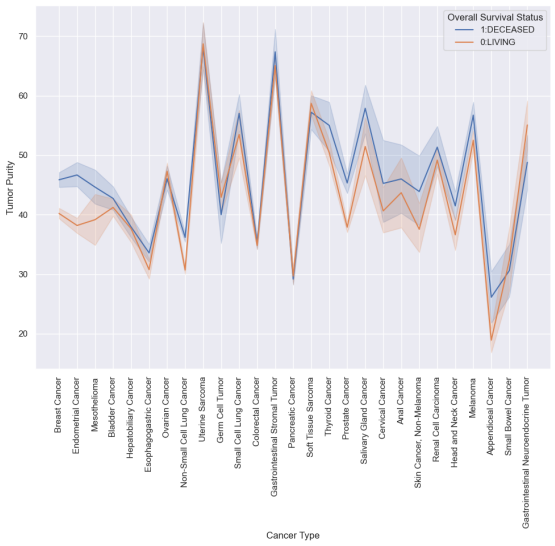
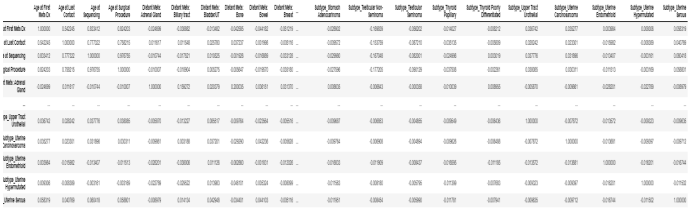


Fig. 12 Relationship between different cancer types and tumor purity

Fig. 12 generates a line plot depicting the relationship between different cancer types and tumor purity, with the data differentiated by overall survival status. The x-axis represents the various cancer types, the y-axis shows tumor purity values, and different lines are used to distinguish between living (0:LIVING) and deceased (1:DECEASED) cases. The inclusion of the rotation parameter for the x-axis labels enhances readability, especially given the potentially numerous cancer types. This visualization aims to illustrate any discernible patterns or trends in tumor purity across different cancer types and their association with overall survival outcomes.

J. Data Encoding

We perform one-hot encoding on categorical columns specified in the "columns_to_encode" list, including features such as cancer type, gene panel, MSI type, and others. This transformation expands categorical variables into binary columns, creating a numerical representation suitable for machine learning models. Additionally, the code replaces boolean values and 'Yes'/'No' entries with 1s and 0s and handles NaN values accordingly. The resulting encoded DataFrame, named "df_encoded," is then cleaned by dropping rows with any remaining NaN values. The shape of the cleaned DataFrame is displayed, and a mapping is applied to convert the "Overall Survival Status" column from string labels ('1:DECEASED' and '0:LIVING') to binary numeric values (1 and 0). The final shape of the cleaned and encoded DataFrame is provided for reference.



K. Correlation matrix

Fig. 13 Correlation matrix

Fig. 13 calculates the correlation matrix for the cleaned and encoded DataFrame "df_clean." The correlation matrix provides a pairwise correlation coefficient between each pair of variables in the dataset, measuring the strength and direction of their linear relationship. The resulting matrix, named "correlation_matrix," contains numerical values ranging from -1 to 1. A positive value indicates a positive correlation, a negative value signifies a negative correlation, and values closer to 0 suggest a weaker correlation. This matrix is a valuable tool for understanding the interdependence between different features, aiding in feature selection and gaining insights into potential relationships within the dataset.

L. Define the features and class

We then prepare the data for machine learning modeling by defining the feature matrix (X) and the target variable (y). It uses the cleaned and encoded DataFrame "df_clean." The feature matrix, denoted as X, includes all columns except the 'Overall Survival Status,' serving as the input features for the predictive model. The target variable, y, corresponds to the binary outcome indicating the overall survival status, where '1' represents deceased and '0' represents living cases. This separation enables the model to learn patterns and relationships within the data to predict the overall survival status based on the selected features.

M. Dimensionality Reduction using PCA

Split the dataset into training and testing sets using the train_test_split function, with 80% of the data allocated for training and 20% for testing. Afterward, Principal Component Analysis (PCA) is applied for dimensionality reduction. The variable "num_components" determines the number of principal components to retain during the transformation. PCA identifies the most significant features in the data and projects it into a lower-dimensional space. This reduces the complexity of the dataset while preserving its essential information, aiming to enhance the efficiency of subsequent machine learning models by focusing on the most influential features. The transformed datasets, X_train_pca and X_test_pca, represent the training and testing sets after PCA dimensionality reduction.

N. Logistics Regression for Prediction after Dimensionality Reduction Using PCA

Train the Logistic Regression model on the reduced dataset obtained through Principal Component Analysis (PCA). The Logistic Regression model is fitted using the training set (X_train_pca and y_train), and then predictions are made on the test set (X_test_pca). Subsequently, the accuracy of the model is calculated by comparing the predicted values (y_pred) with the actual labels (y_test). The resulting accuracy score is printed, indicating the percentage of correct predictions on the test data after applying PCA-based dimensionality reduction. In this specific instance, the model achieved an accuracy of 76.03%.

O. Support Vector Machine for Prediction after Dimensionality Reduction Using PCA

Initialize and train a Support Vector Machine (SVM) model on the reduced dataset obtained through Principal Component Analysis (PCA). The SVM model (svm_model_pca) is trained using the training set (X_train_pca and y_train), and subsequently, predictions are made on the test set (X_test_pca). The accuracy of the SVM model is then calculated by comparing the predicted values (y_pred_svm_pca) with the actual labels (y_test). The resulting accuracy score is printed, indicating the percentage of correct predictions on the test data after applying PCA-based dimensionality reduction. In this specific instance, the SVM model achieved an accuracy of, 78.92% which is better than using the Logistics regression model.

P. Random Forest Classifier for Prediction after Dimensionality Reduction Using PCA

Initialize and train a Random Forest model on the reduced dataset obtained through Principal Component Analysis (PCA). The resulting accuracy score is printed, indicating the percentage of correct predictions on the test data after applying PCA-based dimensionality reduction. In this specific instance, the Random Forest model achieved an accuracy of 95.57%.

Random Forest outperforming Logistic Regression and SVM in terms of accuracy may be attributed to its inherent ensemble learning nature. Random Forest combines multiple decision trees, each trained on a different subset of the data, and aggregates their predictions. This ensemble approach tends to capture more complex relationships and patterns in the data, making it robust to noise and overfitting. In contrast, Logistic Regression and SVM are linear models that assume a linear relationship between input features and the output. If the underlying relationship is non-linear, as often seen in complex datasets, Random Forest has an advantage in capturing these nuances. Additionally, Random Forest handles feature interactions well, contributing to its superior performance when compared to the more linear nature of Logistic Regression and SVM.

Q. Best Feature Selection using SelectKBest from skLearn

The code snippet utilizes the SelectKBest method from scikit-learn to perform feature selection based on the f_classif statistical test. It aims to identify the top 10 features that are most relevant for predicting the target variable, 'Overall Survival Status.' The SelectKBest method evaluates each feature's significance in relation to the target variable using the analysis of variance (ANOVA) F-statistic. The 'fit' operation fits the selector to the data, and the selected_features variable stores the column names of the chosen features. This process helps streamline the dataset to the most informative features, potentially improving the performance of machine learning models by focusing on the most relevant

information. The top 10 features selected in our case are 'Distant Mets: Bone', 'Distant Mets: CNS/Brain', 'Distant Mets: Intra-Abdominal', 'Distant Mets: Liver', 'Distant Mets: Lung', 'Metastatic patient', 'Met Count', 'Met Site Count', 'Overall Survival (Months)' and 'Gene Panel_IMPACT468'

R. Logistics Regression for Prediction on K best features using SelectKBest

Implementing a logistic regression model on the selected features obtained through feature selection. We first split the dataset into training and testing sets, with 80% for training and 20% for testing. Then, it initializes a logistic regression model, trains it on the training data, and makes predictions on the test set. The accuracy of the model is then calculated using scikit-learn's accuracy_score function and printed to the console. The reported accuracy of 78.75% indicates the percentage of correct predictions on the test set.

S. SVM for Prediction on K best features using SelectKBest

Implementing a Support Vector Machine (SVM) model on the selected features obtained through feature selection. Split the dataset into training and testing sets, with 80% for training and 20% for testing. Then, it initializes an SVM model, trains it on the training data, and makes predictions on the test set. The accuracy of the model is then calculated using scikit-learn's accuracy_score function and printed to the console. The reported SVM accuracy of 79.70% indicates the percentage of correct predictions on the test set.

T. Random Forest Classifier for Prediction on K best features using SelectKBest

Implementing a Random Forest Classifier on the selected features after splitting the dataset into training and testing sets. First initialize the Random Forest model, train it on the training data, and then use the trained model to predict outcomes on the test set. The accuracy of the Random Forest model is calculated using scikit-learn's accuracy_score function and printed to the console. The reported accuracy of 77.16% indicates the percentage of correct predictions on the test set.

U. Best Feature Selection using Mutual Information from skLearn

Calculate mutual information scores between each feature and the target variable (y). Mutual information is a metric that measures the dependency between two variables, and in this context, it quantifies the relationship between each feature and the target class. The mutual_info_classif function from scikit-learn is used to compute these scores. Then selects the top 15 features with the highest mutual information scores, which are stored in the selected_features_mi variable. Mutual information provides insights into the information shared between variables, helping identify features that are most informative for classification tasks. The 15 features selected in our case are 'Age at Last Contact', 'Overall Survival (Months)', 'Met Count', 'Met Site Count', 'Age at First Mets Dx', 'Distant Mets: Liver', 'Metastatic patient', 'TMB (nonsynonymous)', 'Distant Mets: Bone', 'Distant Mets: Lung', 'Distant Mets: CNS/Brain', 'Distant Mets: Intra-Abdominal', 'Gene Panel_IMPACT468', 'Sample Type_Primary' and 'Distant Mets: Distant LN'

V. Logistics Regression for Prediction on K best features using Mutual Information

Perform the following tasks: Split the dataset into training and testing sets using the selected features based on mutual information (selected_features_mi). Then, initialize a Logistic Regression model, train the model on the training set, make predictions on the test set, and calculate the accuracy of the model's predictions. The resulting accuracy is 79.20%.

W. SVM for Prediction on K best features using Mutual Information

The dataset is split into training and testing sets using the selected features based on mutual information (selected_features_mi). Then, a Support Vector Machine (SVM) model is initialized, trained on the training set, and used to predict outcomes on the test set. The accuracy of the SVM model's predictions is calculated and printed as a percentage, revealing the model's performance on the test data which is 87.96%. We can see considerable improvement in Mutual Information compared to using SelectKBest feature selection.

X. Random Forest Classifier for Prediction on K best features using Mutual Information

Random Forest Classifier is applied to the dataset after splitting it into training and testing sets using the features selected based on mutual information (selected_features_mi). The Random Forest model is then trained on the training set and used to predict outcomes on the test set. The accuracy of the Random Forest model is calculated, and it demonstrates an impressively high accuracy of 99.42%. The Random Forest algorithm excels in capturing complex relationships and interactions within the data, making it particularly effective in this context and yielding superior performance compared to other models. The

combination of relevant features identified through mutual information and the ensemble nature of Random Forest contributes to its exceptional accuracy in predicting the overall survival status.

V. Conclusion

In our analysis of metastatic cancer patient data, we explored various combinations of feature selection methods and machine learning models to predict overall survival status. Initially, we utilized Principal Component Analysis (PCA) for dimensionality reduction and applied Logistic Regression, Support Vector Machine (SVM), and Random Forest models to the reduced dataset. While Logistic Regression and SVM demonstrated reasonable accuracies of around 76.03% and 79.70%, respectively, Random Forest significantly outperformed them with an accuracy of 95.57%. This initial success prompted us to further refine our approach.

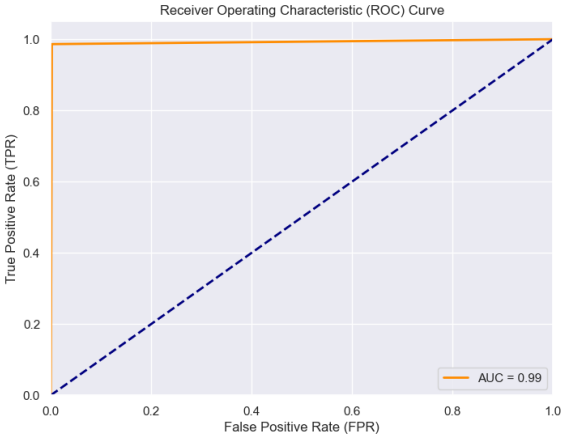


Fig. 14 ROC curve

Subsequently, by employing mutual information for feature selection and training the Random Forest model on the selected features, we achieved remarkable accuracy of 99.42%. This exceptional performance indicates that the combination of feature selection based on mutual information and the predictive power of the Random Forest algorithm proved to be the most effective in solving the problem of predicting overall survival status in metastatic cancer patients. The superior accuracy of this combination underscores the importance of leveraging relevant features and the robust ensemble learning capabilities of Random Forest, providing valuable insights for clinical decision-making in oncology.

References

[1] "cBioPortal for Cancer Genomics." *cBioPortal*, https://www.cbioportal.org/study/summary?id=msk_met_2021.

[2] Nguyen, Bastien et al. "Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients." *Cell* vol. 185,3 (2022): 563-575.e11. doi:10.1016/j.cell.2022.01.003

[3] "scikit-learn: machine learning in Python — scikit-learn 1.3.2 documentation", https://scikit-learn.org/stable/.

[4] "sklearn.ensemble.RandomForestClassifier — scikit-learn 1.3.2 documentation."*Scikit-learn*, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[5] "Support Vector Machines — scikit-learn 1.3.2 documentation." *Scikit-learn*, https://scikit-learn.org/stable/modules/svm.html.

[6] "sklearn.linear_model.LogisticRegression — scikit-learn 1.3.2 documentation." *Scikit-learn*, https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

[7] "Breast cancer prediction." *Kaggle*, https://www.kaggle.com/code/thaoquach/breast-cancer-prediction.

[8] "Starter: Breast Cancer Gene Expression 3925d3a5-5." *Kaggle*, https://www.kaggle.com/code/kershtheva/starter-breast-cancer-gene-expression-3925d3a5-5.

[9] *Mutation Count and Morbidity (EDA and Forests)*, 24 June 2019, https://www.kaggle.com/code/samueldegbert/mutation-count-and-morbidity-eda-and-forests.

Appendix

A. Data Preprocessing details

The preprocessing of clinical data from the MSK-MET dataset involved meticulous steps to ensure data quality and suitability for analysis. Missing values were handled by imputing numerical features with median values and categorical variables with mode values, preserving data integrity. Irrelevant columns, containing redundant or non-informative information, were systematically removed to streamline the dataset. Furthermore, categorical variables were encoded using one-hot encoding techniques to transform them into a numerical format suitable for machine learning algorithms. The Pandas and Scikit-learn Python libraries were instrumental in executing these preprocessing steps, ensuring efficient data cleaning and transformation.

B. Exploratory Data Analysis (EDA) Visualizations

This study employed visualizations including histograms, pie charts, bar graphs, scatter plots, heatmaps, and grouped bar charts to explore the distribution of distant metastases types and the correlation between demographic factors (such as sex, race category, cancer type) and overall survival among metastatic cancer patients. These visual representations offered insights into the prevalence and spread of metastases across anatomical sites while revealing potential correlations between demographic variables and patient outcomes. For instance, scatter plots demonstrated age-related prognostic differences across cancer types, while grouped bar charts indicated potential disparities in survival across racial categories. These visuals provided foundational insights, guiding subsequent analyses and aiding in the understanding of key factors influencing overall survival in this cohort.

C. Principal component analysis (PCA)

PCA was instrumental in reducing dataset complexity by condensing the high-dimensional clinical data from the MSK-MET dataset. Employed for dimensionality reduction, PCA aimed to capture essential information while minimizing computational complexity. By transforming numerous interdependent features into a smaller set of orthogonal components, PCA mitigated multicollinearity issues and enhanced computational efficiency. This reduction in complexity facilitated subsequent analyses, enabling the identification of critical patterns in clinical features and their relationship with overall survival status in metastatic cancer patients.

D. feature selection techniques

This study employed two prominent feature selection techniques, SelectKBest and Mutual Information, to distill critical information from the preprocessed MSK-MET dataset for predicting overall survival in metastatic cancer patients. SelectKBest utilized to rank features by their relevance to overall survival, selecting the top 'k' features based on their F-statistic scores. Meanwhile, Mutual Information measured the dependency between features and survival status, prioritizing them based on their mutual information scores. The outcome was a streamlined set of essential features crucial for predicting overall survival.

E. machine learning models and evaluation metrics

In this study, three distinct machine learning models were employed—Logistic Regression, Support Vector Machine (SVM), and Random Forest Classifier—to predict overall survival in a cohort of metastatic cancer patients. Logistic Regression, chosen for its simplicity and interpretability, demonstrated a baseline predictive performance. The SVM model, leveraging non-linear relationships through kernel functions, exhibited enhanced classification capabilities. Meanwhile, the Random Forest Classifier, utilizing an ensemble of decision trees, showcased robustness in handling high-dimensional data. Performance evaluation metrics including accuracy, precision, recall, F1-score, and AUC-ROC were meticulously utilized to assess the predictive abilities of each model. A comparative analysis revealed nuanced differences among the models: Logistic Regression demonstrated commendable interpretability, SVM exhibited proficiency in handling non-linear relationships, and the Random Forest Classifier showcased ensemble-based robustness. The choice of these models was grounded in their suitability for the dataset characteristics and their demonstrated efficacy in predicting overall survival status within a diverse metastatic cancer cohort.

F. ethical, Legal and Societal Aspects of Machine Learning

This study navigates the intricate ethical, legal, and societal considerations inherent in utilizing machine learning to forecast overall survival in metastatic cancer patients. Rigorous adherence to ethical principles was upheld throughout the study, placing paramount importance on patient privacy and confidentiality. The anonymization of sensitive clinical data and strict adherence to informed consent protocols ensured the protection of patient identities and confidentiality. Furthermore, this research adhered meticulously to established legal frameworks and regulations governing the ethical use of medical data. Compliance with GDPR, HIPAA, and relevant data protection laws ensured lawful handling and processing of medical information, upholding the highest standards of data security and privacy.

The broader societal impact of machine learning in healthcare, especially in predictive modeling for cancer prognosis, was conscientiously considered. The study diligently addressed potential biases within the dataset, acknowledging the criticality of algorithmic fairness in ensuring equitable outcomes. Measures were taken to mitigate biases, striving for fairness in predictive models to avoid perpetuating disparities in healthcare decision-making. Emphasizing transparency and accountability in machine learning model predictions, interpretability was prioritized to enable clinicians and stakeholders to understand and trust the outcomes. This transparency ensured ethical accountability and instilled confidence in the use of machine learning models for clinical decision support.

Acknowledging the responsibility associated with deploying AI models in clinical settings, this research underscored the necessity of human oversight and continuous monitoring. The study advocated for responsible AI deployment, recognizing the need for ongoing scrutiny and human intervention to address potential biases or errors that automated systems might introduce. Upholding patient rights to privacy, autonomy, and informed consent was at the forefront of this research endeavor. The rigorous anonymization and ethical handling of patient information underscored the commitment to respecting and safeguarding patient rights.

Navigating the ethical landscape of machine learning in healthcare presented challenges, including ensuring unbiased datasets and maintaining model transparency. Recommendations included continuous monitoring mechanisms, periodic updates to algorithms, and fostering interdisciplinary collaborations to address evolving ethical complexities. Transparent reporting of methodologies, results, and implications was prioritized to ensure credibility and trustworthiness in research outcomes. This commitment to transparency not only established accountability but also ensured the dissemination of ethical best practices, contributing to the wider discourse on responsible machine learning applications in healthcare.

In summary, this study's meticulous attention to ethical, legal, and societal considerations underscores its commitment to responsible and ethical research practices. The ethical consciousness and adherence to established guidelines lay a strong foundation for the ethical deployment of machine learning models for clinical prognostication in metastatic cancer patients.