**Understanding the Impact of Class Imbalance on Clinical Endpoint Prediction in Neuroblastoma with Different Supervised Machine Learning Methods**

**Isil Erbasol Serbes (22.10.2023)**

**ABSTRACT**

Neuroblastoma, a predominant pediatric cancer, poses formidable challenges in patient prognosis and treatment. In this study, I harnessed neuroblastoma microarray data in conjunction with essential clinical features, such as MYCN status and clinical genetic subgroups, to predict some clinical endpoints: INSS.Stage, HighRisk, Progression, and DeathFromDisease. Employing three distinct machine learning algorithms—Random Forest, Radial Kernel SVM, and Logistic Regression—I developed and evaluated predictive models. Both the complete feature matrix and a refined feature selection were considered in the analysis. The study resulted in six predictive models for each clinical endpoint, totalling 24 models. Model performance was rigorously assessed through various metrics, including confusion matrices, ROC-AUC, and PR-AUC. From the model evaluations, I observed that the HighRisk clinical endpoint exhibited outstanding model performance, with the potential for accurate prediction. However, for other clinical endpoints characterized by imbalanced subclasses, the prediction results were less promising, particularly for minority classes. The majority classes mostly consistently achieved commendable predictive accuracy. These findings emphasize the critical importance of balanced class distribution for robust clinical endpoint predictions and encourage further exploration of strategies to enhance model performance for minority classes.

**INTRODUCTION**

Neuroblastoma, a predominantly pediatric cancer, presents a multifaceted challenge in the realms of patient prognosis and treatment (Gains et al., 2012, Matthay et al., 2016, Morgenstern et al., 2013). Its intricate nature calls for innovative strategies that can enhance diagnostic accuracy and elevate the quality of patient care.

The application of machine learning techniques in cancer diagnosis holds great promise for revolutionizing the field of oncology. By leveraging advanced algorithms, these approaches can analyze complex data patterns to improve early detection and tailored treatment strategies, ultimately contributing to more accurate diagnoses and better patient outcomes (Swanson et al., 2023, Saberi-Karimian et al., 2021, Kononenko, 2001, Iqbal et al., 2021). Among different supervised machine learning algorithms, Random Forest, Support Vector Machine (SVM), and Logistic Regression stand out as prominent machine learning algorithms utilized in clinical diagnosis, particularly in the field of cancer (Qian et al., 2021, Huang et al., 2018, Zangmo and Tiensuwan, 2018, Khairunnahar et al., 2019, Acharjee et al., 2020, Dai et al., 2018). These algorithms are instrumental in deciphering intricate data patterns, aiding in early detection, and optimizing personalized treatment strategies, thereby advancing the accuracy of medical diagnoses, and ultimately improving patient outcomes.

Microarray-based gene expression profiling has emerged as a formidable tool in cancer research allowing researchers to delve deep into the molecular underpinnings of the disease (Ohira et al., 2005a, Ohira et al., 2005b). By simultaneously examining the activity of thousands of genes, this technology enables the comprehensive exploration of gene signatures, uncovering intricate relationships between genetic variations and cancer development. This wealth of genomic information, when combined with advanced machine learning algorithms like Random Forest, Support Vector Machine (SVM), and Logistic Regression, equips oncologists with cost-effective tools to better understand the complex molecular landscape of cancer. Consequently, it offers opportunities for more precise, affordable, and personalized treatment strategies, ultimately leading to improved diagnostic accuracy, enhanced patient care, and better overall outcomes (Shamai et al., 2019, Merseburger et al., 2006).

In this study, I aimed to evaluate the effectiveness of machine learning models in predicting crucial clinical endpoints (INSS.Stage, HighRisk, Progression, and DeathFromDisease) for neuroblastoma patients by using the integrated analysis of microarray data and clinical information. To accomplish this, I harnessed the capabilities of three distinct machine learning algorithms: Random Forest, Radial Kernel SVM, and Logistic Regression. My approach was comprehensive, considering both the complete feature matrix and a carefully refined feature selection. This holistic analysis yielded a total of six predictive models for each of the clinical endpoints, resulting in a grand total of 24 models. From my extensive model evaluations, one key observation stood out: the HighRisk clinical endpoint demonstrated remarkable model performance, offering the potential for accurate predictions. However, for the remaining clinical endpoints, characterized by imbalanced subclasses, the results were less promising, particularly for minority classes. The majority classes consistently exhibited commendable predictive accuracy. These findings underscore the critical significance of balanced class distribution in achieving robust clinical endpoint predictions. They also serve as an encouragement to further explore strategies that can enhance model performance, particularly for minority classes, and contribute to the advancement of clinical diagnosis and patient care in the context of neuroblastoma. These efforts contribute to the advancement of clinical diagnosis and patient care in the context of neuroblastoma.

**METHODOLOGY**

**Program and software**

Unless mentioned otherwise, using the Python machine learning package, scikit-learn 1.3.1, data preprocessing and all algorithms were employed with various parameters. For visualization purposes, Python's Matplotlib 3.8.0 and Seaborn 0.13.0 packages were utilized.

**Dataset**

In this study, a previously curated microarray dataset along with comprehensive clinical metadata of neuroblastoma patients from Zhang et al. study was used (Zhang et al., 2015). The clinical metadata encompasses clinical features, including Gender, Age, clinico-genetic subgroup, and MYCN status, in conjunction with critical clinical endpoints, namely INSS.Stage, High Risk, Progression, and DeathFromDisease, derived from a cohort of 373 neuroblastoma patients (Figure1).

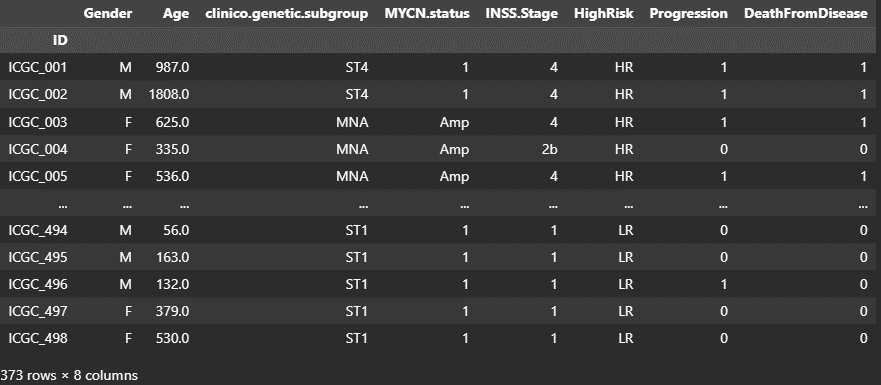


Figure : General overview of Clinical metadata

The "Gender" feature distinguishes between female and male patients and, "Age" represents the age of patients at the time of diagnosis, measured in days. Approximately 90% of these tumors occur in children under the age of 10, and the typical age at which neuroblastoma is diagnosed is around 18 months (Matthay et al., 2016). The "clinico-genetic subgroup" variable stratifies patients into four distinct subgroups: MTA (MYCN Amplified, any INSS Stage, ≤18 and >18 months), ST1 (No MYCN Amplified, Stage 1, ≤18 and >18 months), ST4 (No MYCN Amplified, Stage 4, ≤18 and >18 months) (Figure2), and ST4S (No MYCN Amplified, Stage 4, ≤18 months). And, The "MYCN.status" clinical feature provides a binary classification, indicating whether MYCN gene, a genetic marker for neuroblastoma (Otte et al., 2021) is amplified (Amp) or not (1).



Figure : clinico.genetic.subgroups (Zhang et al., 2015)

Clinical endpoint "INSS Stage" is a multi-category variable representing the International Neuroblastoma Staging System with seven subgroups (1, 2, 2a, 2b, 3, 4, and 4S) where disease severity as following: Stage1-2: Mild, Stage3 -Stage4S: Moderate and Stage4: Severe (Castel et al., 1999). The "HighRisk" allows clinical categorization into High Risk (HR) and Low Risk (LR) neuroblastoma. "Progression" captures the occurrence of tumor progression events, denoting 'yes' as 1 and 'no' as 0, shedding light on disease dynamics. "DeathFromDisease" similarly records patient outcomes, distinguishing between 'yes' (1) and 'no' (0) for instances of death from the disease.

Lastly, the microarray dataset encompasses expression values of 44,708 genes, obtained from 498 neuroblastoma patients.

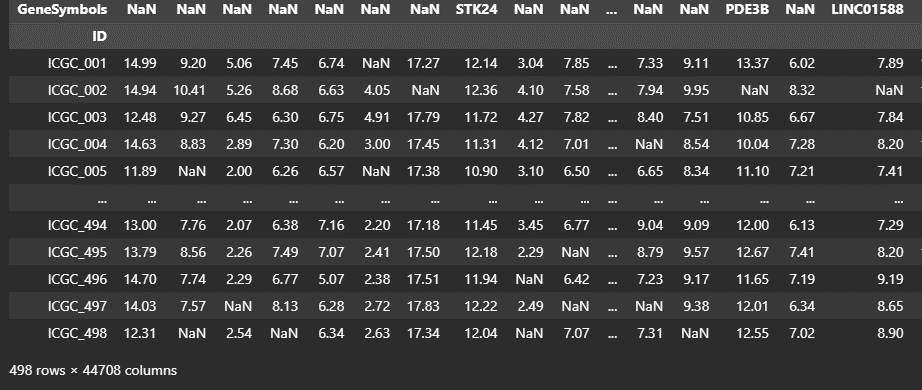


Figure :Figure 1: General overview of microarray data

The microarray dataset reveals that there are 498 samples (probes) and 44708 genes in the dataset.

**Explanatory Data Analysis**

Exploratory Data Analysis (EDA) involves several components. Descriptive Analysis was used to conduct a thorough examination of the dataset's summary statistics to gain insights into central tendencies, missing values, dispersions, and the overall characteristics of the variables. Feature and Clinical Endpoint Visualization were used to explore relationships and outliers within the dataset, with bar-box plots used for visualization. Cross-Tabulation was employed to investigate relationships between different variables, particularly for exploring associations between clinical endpoints and various features.

**Data Pre-processing step**

In the data pre-processing step, I made several adjustments to ensure the consistency and usability of the dataset. First, I changed the MYCN subclass name "1" to "NoAmp" for consistency, aligning the naming convention across the dataset. Additionally, I assigned INSS.Stage 2a and 2b subcategories into a single category named "Stage2" to simplify and standardize the staging system. I removed any additional tags from microarray data sample names, ensuring clean and uniform labels. Categorical features and INSS.Stage clinical endpoints were one-hot encoded to transform them into a suitable format for machine learning algorithms. Next, I merged the clinical and microarray data based on common samples, resulting in a new dataset with 373 samples. I removed genes with no expression values, optimizing the dataset for analysis. To facilitate model training and evaluation, I split the dataset into training and test sets, reserving 20% for testing. Missing gene expression values in the microarray data were imputed using a median strategy to handle gaps in the dataset. Finally, I applied data normalization to continuous variables: age and gene expressions using standard scaling. This step ensured that all features were on a common scale, preventing any individual feature from dominating the analysis.

**Unsupervised Machine Learning**

To identify natural patterns and outliers ultimately leading to a more comprehensive understanding of the underlying structure in the dataset. I performed unsupervised machine learning methods. I conducted several key techniques, including Principal Component Analysis (PCA) and K-means clustering. Additionally, I explored the combination of K-means and PCA to gain deeper insights into the data.

**Feature importance**

I conducted an in-depth analysis of feature importance for classification tasks, aiming to understand the critical features that influence the model's decision-making process. To achieve this, I employed an ensemble learning technique: Random Forest to capture feature importance. The approach was structured as follows.I began by initializing a Random Forest Classifier with 100 decision trees and a fixed random seed for reproducibility. The classifier was configured to handle the classification task efficiently. I then adopted a Stratified K-Fold cross-validation strategy with five splits, ensuring that each fold preserved the original class distribution. This method is particularly valuable when working with imbalanced datasets, as it guarantees that each class is represented fairly in both the training and testing sets. For each cross-validation split, I fitted the Random Forest Classifier to the training data, extracting feature importances from the trained model. By calculating the mean feature importances and sorting them in descending order, I obtained a list of common critical features. Finally, I visualized the top common critical features using a horizontal bar chart.

**Feature selection**

I employed a feature selection methodology to identify a subset of informative features for classification. I utilized a Stratified K-Fold cross-validation approach with five splits to ensure robustness and unbiased evaluation. To select the relevant features, I used SelectKBest and SelectPercentile, which are feature selection methods that evaluate feature importance based on statistical metrics. The top 300 features were chosen as they consistently demonstrated strong discriminative power across different folds. For each split in the cross-validation, I compared the features selected by both SelectKBest and SelectPercentile and identified the common features shared by both methods. These common features were regarded as the most robust and informative for the classification tasks. I trained a Random Forest Classifier on each split using the selected features and computed the accuracy to evaluate the classification performance. The final step involved combining the results from all folds to identify the intersection of common features, resulting in a set of features that exhibited consistent importance in all cross-validation iterations. This set of "final common features" represents a subset of attributes that are likely to be the most influential for the classification problems. These features further used during model development.

**Hyperparameter tuning and stratified cross validation**

For each clinical endpoint prediction, three different supervised machine learning algorithms namely Random Forest, Radial Kernel Support Vector Machine and Logistic Regression along with two different feature matrices, one with selected features and one with whole features were used. For each model, hyperparameter tuning were either with Grid search (For Random Forest and Logistic Regression) or Optuna (Akiba et al., 2019) (Radial kernel SVM) performed. Afterwards, three-fold Stratified cross validation with best hyperparameters for each model performed and model results were evaluated with different metrices.

**Model evaluation**

In the comprehensive evaluation of all 24 models, I utilized a diverse array of evaluation metrics to understand of their performance. The confusion matrix was used visualize true positive, true negative, false positive, and false negative predictions, vital in understanding the model's ability to correctly classify instances. Precision, a measure of proportion of true positive predictions among all positive predictions, was employed highlighting the model's ability to make accurate positive classifications. Accuracy, which quantifies the overall correctness of predictions, was utilized the view of model performance. Recall also known as sensitivity or true positive rate was used to see the model's capacity to identify all actual positive instances, guarding against missed positive cases. F1-score, a harmonious balance of precision and recall, was employed to gain insight of model performance in presence of imbalanced data. In addition to these matrices, ROC-AUC (Receiver Operating Characteristic - Area Under the Curve, crucial for assessing the model's discrimination ability in binary classification tasks and PR-AUC (Precision-Recall Area Under the Curve), giving insights into precision and recall trade-offs, especially when dealing with imbalanced datasets were used. Each of these metrics enabled to make informed decisions about prediction ability for clinical endpoints.

**RESULTS**

**1. Explanatory Data Analysis**

**1.1 Clinical Data**

**Descriptive analysis for clinical data**

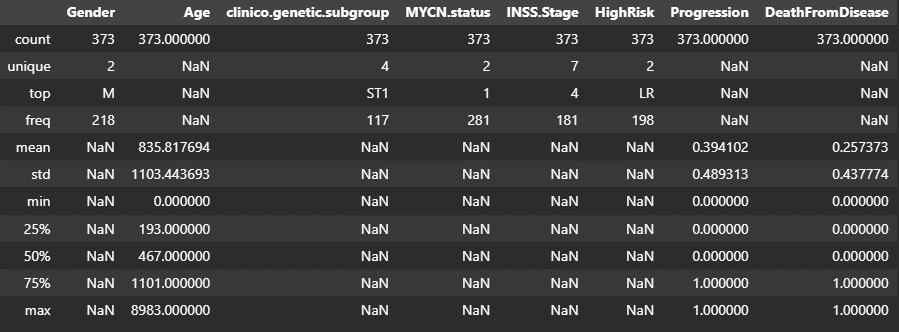


Figure :Descriptive statistics of clinical data

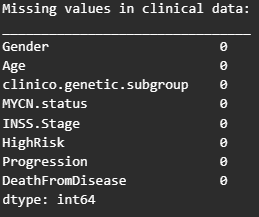
The descriptive analysis of the clinical data reveals several key aspects (Figure4). Notably, there are no missing values in the dataset, indicating its completeness (Figure5). In terms of gender distribution, there is a slight imbalance, with 218 male samples out of a total of 373. An intriguing finding is the presence of a maximum age value of 8983 days, which corresponds to approximately 24.58 years. Given that neuroblastoma primarily affects infants and young children, these older individuals in the dataset might be considered natural or artificial outliers and warrant further investigation (Matthay et al., 2016). In the clinic.genetic.subgroup, there are four categories, with ST1 being the most prevalent. The MYCN.status features two categories, with '1' being the more common category, represent 'no amplification'. The INSS.Stage category seemingly contains seven categories, despite INSS itself having only six (Society). Category '4' is the most frequent which indicates severe form of the disease. High Risk has two unique categories (HR-LR), with 'LR' being the dominant one. The Progression feature contains two unique categories (0-1) representing no progression and progression, while 'DeathFromDisease' features two categories (0-1) signifying individuals who are alive and those who have succumbed to the disease.

Figure : Missing values in clinical data

**Visual inspection of the clinical features**

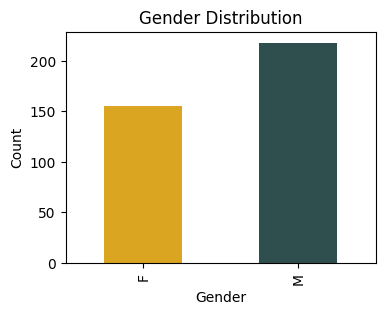


Figure :Gender Clinical Feature

**Gender:** Visualizing the clinical features in the dataset reveals a gender distribution where there are fewer females (152 samples) compared to males (218 samples) (Figure6).

**Age:** In the process of analyzing the age feature, it became evident that there were outliers within this variable (Figure7). As Matthay et al. (2016) pointed out, approximately 90% of tumors manifest in children under the age of 10. With this knowledge in mind, we contemplated whether these age outliers might be inherent to the dataset. Since these outliers have the potential to introduce bias in the data, it was imperative to address them. Upon closer examination, it was observed that these outliers primarily consisted of individuals above the age of 15 and were limited to only three samples. As a result, the decision was made to exclude these outliers from the dataset when the age variable is considered during the feature selection process.

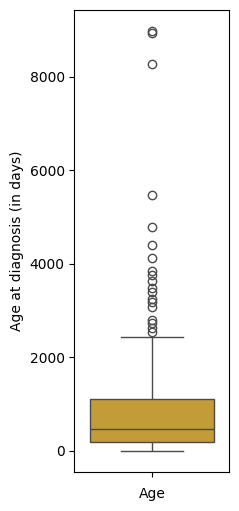


Figure :Age Clinical Feature

**Clinico genetic subgroup:** The Clinico-genetic subgroups within the dataset reveal an inherent class imbalance, with ST4S exhibiting the lowest count among the subgroups, and MNA and ST4 having lower counts compared to ST1 and ST4 (Figure8). Recognizing the potential impact of class imbalance on model performance, if the Clinico genetic subgroup variable is considered during feature selection, employing imbalanced sampling techniques such as oversampling can be required to improve model performance.

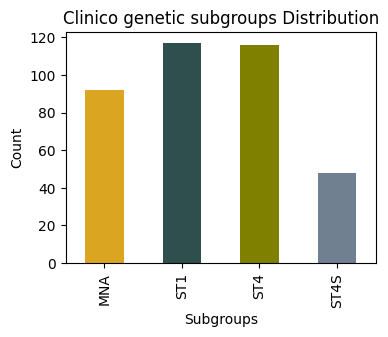


Figure :clinico.genetic.subgroups clinical feature

Furthermore, it’s essential to note that the clinico-genetic subgroup is intricately related to age, INSS.Stage, and MYCN.status, making it unsuitable as a standalone feature for INSS.Stage prediction. It cannot also be used together with the Age and/or MYCN.status clinical features due to its relation.

**MYCN.status**: The MYCN.status feature exhibits a class imbalance with 1 (no Amp) having a higher count (281) compared to Amp (92) (Figure9). Again, if the MYCN.status variable is considered during feature selection, employing imbalanced sampling techniques such as oversampling can be required to improve model performance.Additionally, due to the consistence, 1 subgroup was renamed as 'NoAmp’. This change serves to provide more informative category names, facilitating a better understanding of the data.

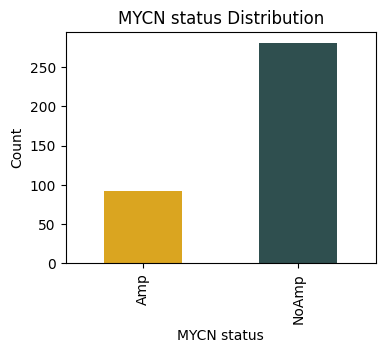


Figure :MYCN.status clinical feature

**Inspection of the clinical endpoints**

**INSS.Stage:** INSS.Stage shows different subcaterogies indicating a multiclass classification problem (Figure 11). In addition, upon closer inspection, a significant imbalance was observed among various subcategories, which is likely to impact the model's predictive performance regarding this clinical endpoint. However, in this study, the emphasis will not be on employing imbalanced sampling methods, but rather on enhancing the model's overall performance. Moreover, it was seen that for Stage2, there are different categories (Figure 10). In general, Stage 2 is either referred as Stage 2 or by its subgroups; Stage 2a and Stage2b. Because Zhang et al in their study, referred this stage as Stage2 only (Zhang et al., 2015) (Figure10), and all the data was not a duplication but unique, therefore all of the subgroups of Stage2 was renamed as Stage2.

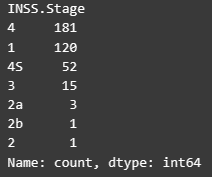


Figure : Number of samples in different INSS.Stages subgroups

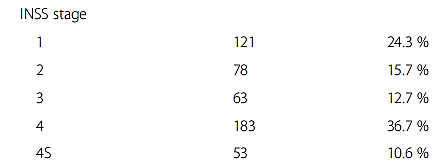


Figure : Original reference to Stage 2 in Zhang et al paper (Zhang et al., 2015)

**High.Risk:** The HighRisk feature demonstrates a balanced distribution among its subcategories (Figure12). Specifically, there are 198 samples categorized as 'LR' (Low Risk) and 175 samples categorized as 'HR' (High Risk). This balance between the subcategories ensures that no single category dominates the dataset, which can lead to more equitable and reliable model training and predictions for this particular feature.

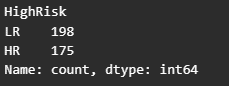


Figure HighRisk subgroups

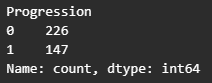
**Progression:** Similar to INSS.Stages, Progression is another clinical endpoints which subclasses shows a clear imbalance distribution where there are 226 samples categorized as '0' (indicating 'no progression') and 147 samples categorized as '1' (signifying 'progression') (Figure13). These imbalanced subclasses can affect the model performance and accuracy for predicting the subclasses.

Figure : Progression subgroups

**DeathFromDisease:** The DeathFromDisease feature displays a noticeable imbalance among its subcategories. Specifically, there are 277 samples categorized as '0' (indicating 'alive') and 96 samples categorized as '1' (representing 'death' due to the disease) (Figure14). This imbalance highlights the need for careful consideration and potential application of imbalanced data handling techniques when working with this feature in order to achieve better model performance. However, this application is out of scope for this study.

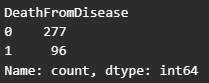


Figure :DeathFromDisease subgroups

**Exploring Clinical Data Through Cross-Tabulation**

Cross-tabulation is a powerful analytical tool that allows us to uncover relationships between categorical variables, test hypotheses, and make informed decisions based on data. Here Cross-Tabulation was employed to investigate pairwise relationships between clinical endpoints and clinical features

**Gender vs Clinical endpoints**

**Gender vs INSS Stage:** In the cross-tabulation of Gender and INSS Stage, intriguing patterns emerge (Figure15). INSS Stage 2b is exclusively observed in females, while INSS Stage 2 is predominantly observed in males, suggesting gender-specific tendencies in disease staging. Furthermore, there is a notable gender disparity in Stages 4 and 4S, with a higher representation of males. In contrast, INSS Stage 2, in general, is underrepresented for both genders, implying its relative rarity in our dataset. If Gender feature is selected for INSS.Stages, there is a required of being mindful of potential imbalance representations of gender in different INSS.Stages (Feature bias) .

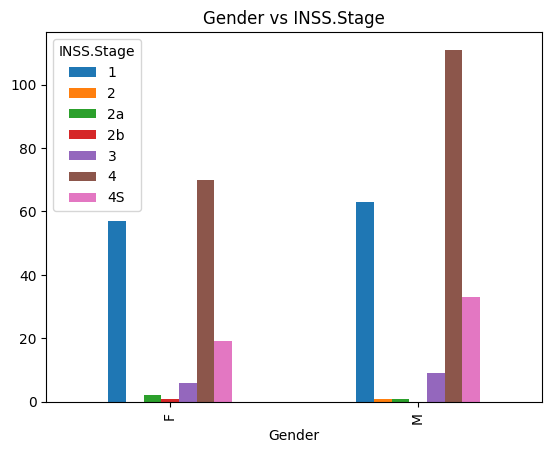


Figure :Cross tablulation between Gender and INSS.Stages

**Gender vs HighRisk:** The data shows that there are more male samples categorized as 'HR' (High Risk) compared to females (Figure16). This suggests that the Gender feature may introduce a slight bias in model learning, especially when predicting HighRisk. Gender appears to be associated with HighRisk status, and this should be taken into account when building the predictive model.

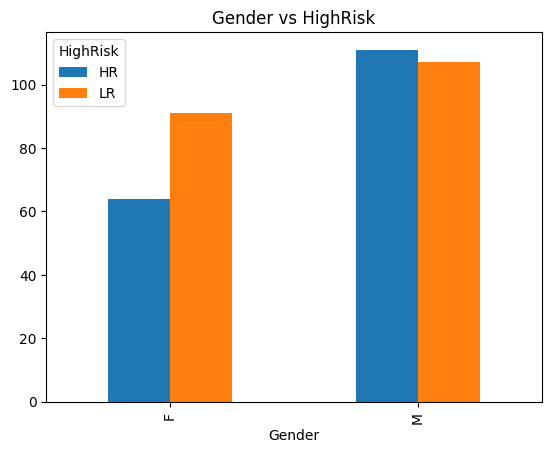


Figure :Cross tablulation between Gender and HighRisk

**Gender vs Progression:** It is apparent that there are different distributions of Progression between genders (Figure17). In the '0' category (indicating 'no progression'), both genders have a substantial number of cases, but the male group is larger. Similarly, in the '1' category (signifying 'progression'), both genders have cases, but the male group is also larger. This suggests that Gender might also be a bias feature in predicting Progression.

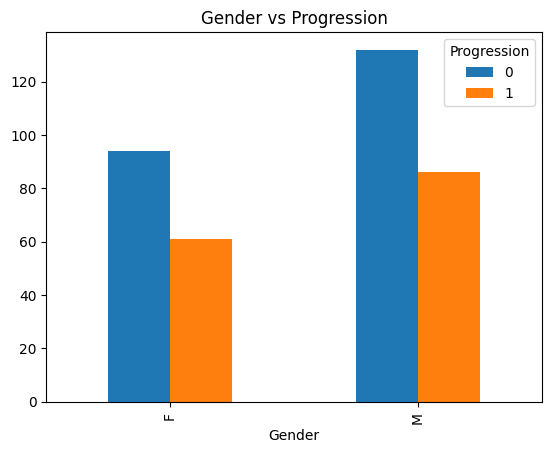


Figure :Cross tablulation between Gender and Progression

**Gender vs DeathFromDisease:** There is a clear difference in the distribution of DeathFromDisease between genders (Figure 18). In the '0' category (indicating 'alive'), both genders have cases, with males having a larger number. In the '1' category (representing 'death' due to the disease), both genders also have cases, but again, males have a larger number. This implies that Gender may also be bias feature in predicting DeathFromDisease and this should be taken into account when developing the predictive model. To ensure the model does not only consider these gender-related patterns during training.

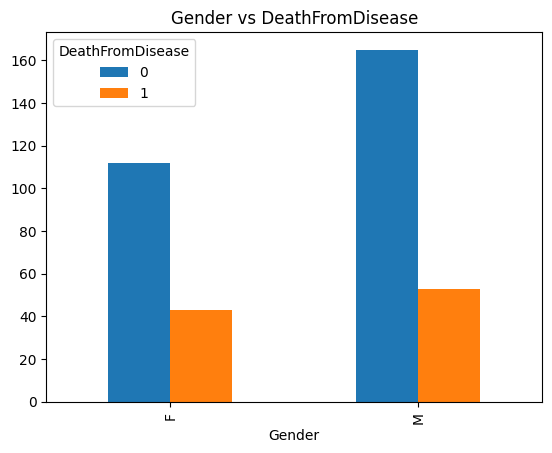


Figure :Cross tablulation between Gender and DeathFromDisease

In summary, Gender exhibits noteworthy feature bias among clinical endpoints expect for HighRisk which is not dominant. For INSS Stage, distinct gender-specific patterns emerge, including exclusive observations of INSS Stage 2b in females and a predominant occurrence of INSS Stage 2 in males. Furthermore, both Progression and Death From Disease are influenced by Gender, with males having larger representations in certain categories, suggesting a need to consider Gender as bias feature.

**Age vs Clinical endpoints**

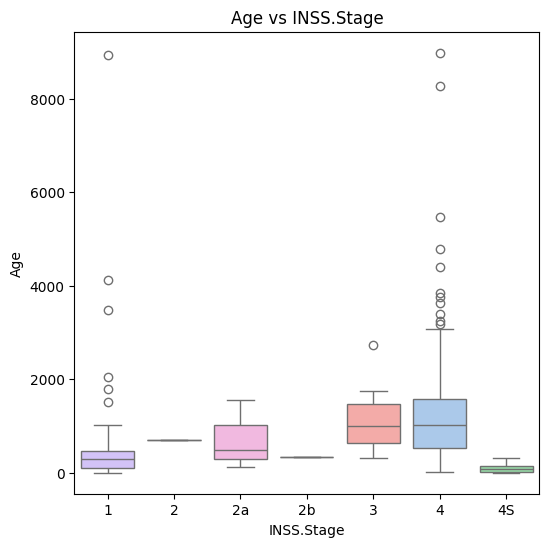


Figure :Cross tabulation between Age and INSS.Stage

**Age vs INSS.Stages:** The variability in age values across different INSS Stages, with a notable exception in Stages 3 and 4, raises a critical concern (Figure20). This relationship between age and INSS Stage has the potential to introduce bias when incorporating age as a feature in predictive modeling or statistical analysis. If a strong correlation exists between age and INSS stage, the utilization of age as a predictor may inadvertently lead the model to base predictions on age rather than accurately considering the genuine underlying determinants influencing INSS stage.

**Age vs HighRisk:** In terms of Age, it appears that individuals categorized as High Risk tend to be older compared to those in the Low Risk category (Figure19). This relationship between age and HighRisk is a concern for introducing bias when incorporating age as a feature in predictive modeling or statistical analysis.

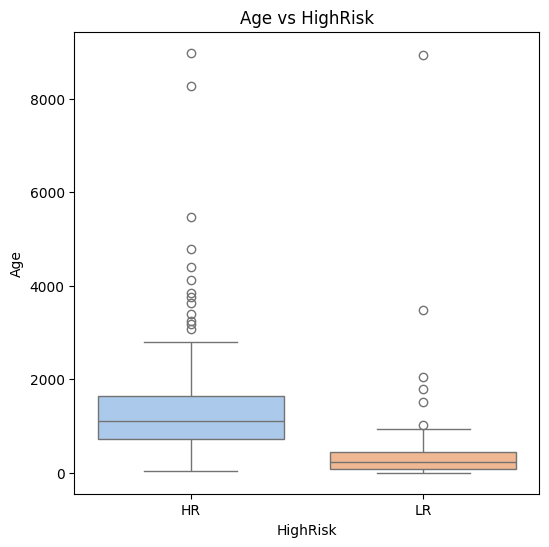


Figure :Cross tabulation between Age and HighRisk

**Age vs Progression:** In the context of Progression, it's noticeable that individuals who experience progression have a higher age at diagnosis compared to those who do not progress (Figure21). This age difference between the two groups may hold bias again in predictive modeling.

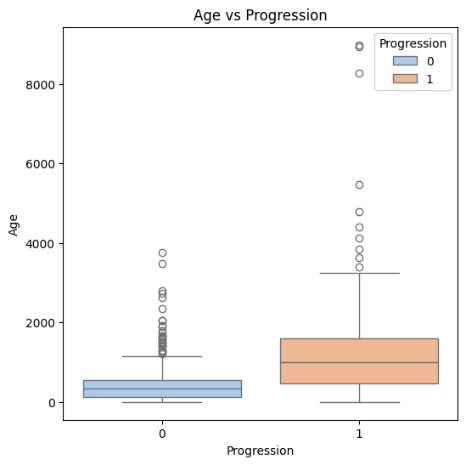


Figure :Cross tabulation between Age and Progression

**Age vs Death From Disease:** When examining Age in the context of DeathFromDisease, it becomes evident that individuals who unfortunately succumb to the disease tend to have a higher age at diagnosis compared to those who do not (Figure22) . This age difference between the two groups suggests a potential relationship between age and the risk of death from the disease, which is an essential consideration in the analysis and predictive modeling.

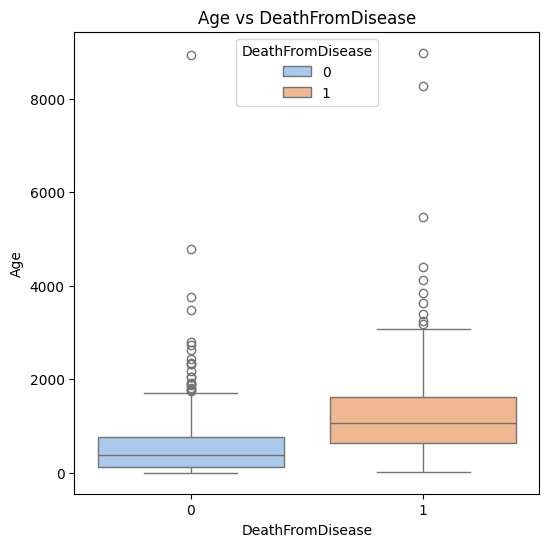


Figure :Cross tabulation between Age and DeathFromDisease

**clinico.genetic.subgroup vs Clinical endpoints**

**clinico genetic subgroup vs INSS Stage**

The reference paper (Zhang et al., 2015) underscores the formulation of the clinico-genetic subgroup, which is intricately based on INSS Stage, Age, and MYCN status. This close connection between the clinico-genetic subgroup and these critical clinical factors raises concerns about its suitability as a predictive feature. Given its inherent bias and strong ties to INSS.Stages, Age, and MYCN.status, it becomes apparent that the clinico-genetic subgroup may not be suitable for predicting INSS.Stages and combinational use of Age and MYCN.status in the feature matrix.

**clinico.genetic subgroup vs HighRisk:** The observation of the clinico-genetic subgroup as a potential bias feature for High Risk is significant (Figure23). Specifically, it's noted that High Risk (HR) is exclusively observed in the MNA subgroup, while Low Risk (LR) is exclusively observed in the ST1 and ST4S subgroups. This strong association between the clinic.genetic subgroup and HighRisk further underscores the need for careful consideration when using clinic.genetic subgroup as a predictor for High Risk in the analysis.

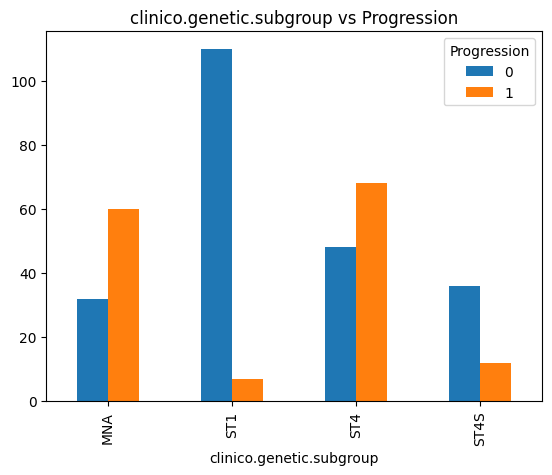


Figure :Cross tabulation between clinico.genetic.subgroup and Progression

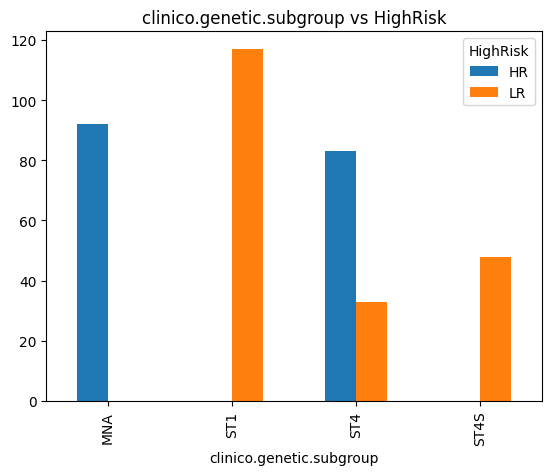


Figure :Cross tabulation between clinico.genetic.subgroup and HighRisk

**clinico.genetic.subgroup vs Progression:** The cross-tabulation of Progression and clinic.genetic subgroup indicates that there is a notable association between clinic.genetic subgroup and Progression (Figure24). Specifically, the MNA and ST4 subgroups exhibit similar results in terms of Progression. This observation suggests that clinico.genetic subgroup is also a potentially biased feature for Progression, reinforcing the need to carefully consider its impact when using it as a predictor in Progression analysis.

**clinico.genetic.subgroup vs DeathFromDisease:** The cross-tabulation of DeathFromDisease and clinico-genetic subgroup highlights another important observation: clinico-genetic subgroup demonstrates a clear association with DeathFromDisease (Figure25). Notably, the ST1 subgroup exhibits a pattern of zero cases in the '1' category (indicating death due to the disease), while other subgroups have cases in both categories. This signifies that clinico-genetic subgroup is also a potentially biased feature for predicting DeathFromDisease.

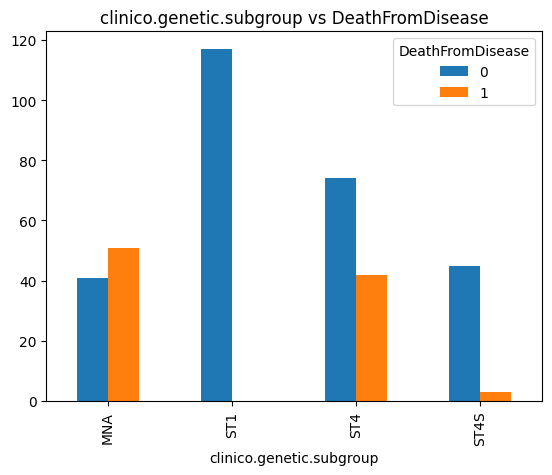


Figure :Cross tabulation between clinico.genetic.subgroup and DeathFromDisease

In summary, the clinic.genetic.subgroup, as emphasized by the reference paper, is intricately linked to INSS Stage, Age, and MYCN status (Zhang et al., 2015), posing challenges for its suitability as a predictive feature for INSS.Stage. Furthermore, the association of clinic.genetic.subgroup with HighRisk, Progression, and Death From Disease implies potential bias in predictive modeling for these clinical endpoints, necessitating careful consideration and management to ensure accurate predictions while accounting for clinic.genetic.subgroup influences.

**MYCN.status vs Clinical endpoints**

**MYCN.status vs INSS.stage:** MYCN.status appears to introduce bias in predicting INSS.Stage (Figure26). It's noteworthy that while one might expect MYCN amplification (Amp) to be associated with greater severity, No amplification (NoAmp) is observed across both INSS Stage 1 and Stage 4, whereas Amp is primarily seen in Stage 4. This unexpected distribution suggests that MYCN.status may not align as strongly with disease severity as anticipated and warrants careful consideration when using it as a feature in INSS.Stage prediction. Its influence on the model's performance should be rigorously assessed to ensure accurate predictions while accounting for these patterns.

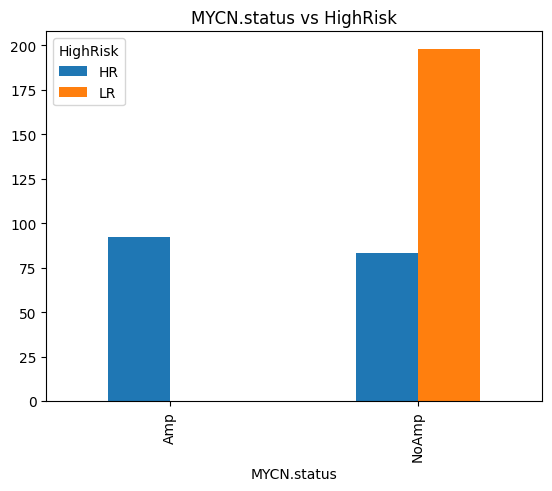


Figure :Cross tabulation between MYCN.status vs HighRisk

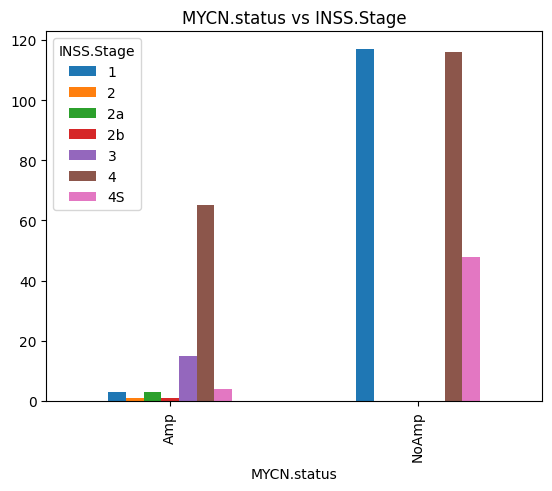


Figure :Cross tabulation between MYCN.status vs INSS.stage

**MYCN.status vs HighRisk:** MYCN.status is another a bias feature for predicting HighRisk (Figure27). Notably, Amplified MYCN (Amp) is associated with HighRisk (HR), and no LowRisk (LR) patients are observed in this category. This suggests a strong correlation between MYCN amplification and HighRisk. The clear association between MYCN.status and HighRisk should be considered in predictive modeling for HighRisk, and the potential bias should be addressed to ensure accurate predictions while accounting for this correlation.

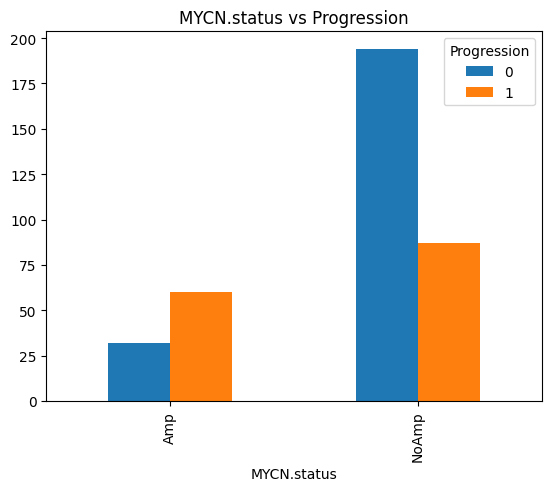


Figure :Cross tabulation between MYCN.status vs Progression

**MYCN.status vs Progression:** MYCN.status also introduces bias in predicting Progression (Figure29). Notably, when there is No amplification (NoAmp), there are more patients with no progression (0), whereas Amplified MYCN (Amp) is associated with a higher proportion of patients experiencing progression (1). This indicates a clear correlation between MYCN amplification and Progression, and this bias should be carefully managed when using MYCN.status as a feature in Progression prediction.

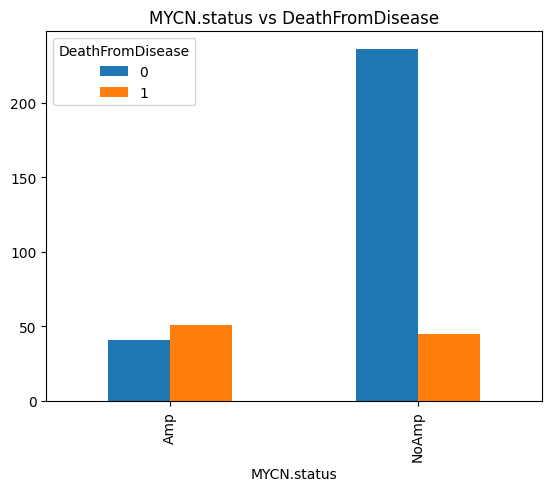


Figure :Cross tabulation between MYCN.status vs DeathFromDisease

**MYCN.status vs DeathFromDisease:** MYCN.status also introduces bias in predicting DeathFromDisease (Figure28). When there is No amplification (NoAmp), there is a higher proportion of patients marked as 'alive' (0), whereas Amplified MYCN (Amp) is associated with both 'death' (1) and 'alive' (0) patients. This interesting distribution implies a complex relationship between MYCN amplification and DeathFromDisease, and it should be carefully managed when using MYCN.status as a feature in DeathFromDisease prediction.

In summary, MYCN.status introduces bias in predicting clinical endpoints, including While Amplified MYCN (Amp) might be expected to align with greater disease severity, its distribution across categories challenges this assumption. The strong associations observed suggest a need for careful consideration of MYCN.status as a predictive feature in these clinical endpoints, with the potential bias warranting thorough assessment to ensure accurate predictions while accounting for these correlations.

**1.2 Microarray data**

**Descriptive analysis for microarray data**

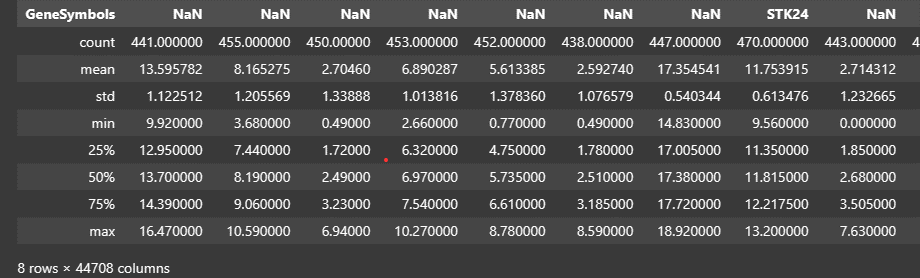


Figure :Descriptive statistics of microarray data

The descriptive analysis of the microarray data reveals several key aspects (Figure30). Firstly, it shows the presence of unannotated gene names and missing gene expression values (Figure 30-31), indicated by varying count values, necessitating imputation. Furthermore, an examination of the median values suggests variation in gene expression, highlighting the need for normalization. In addition, some probe names have additional tags which can cause some issues during merging with clinical datasets. Therefore, these additional tag attachments must be removed from the sample names. Furthermore, there is a discrepancy in the sample numbers when compared to the clinical data. Therefore, it is necessary to merge the common samples from both datasets for further analysis.

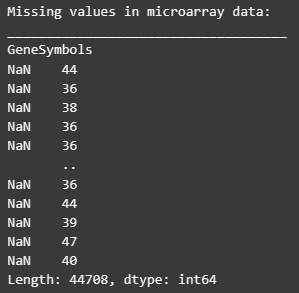


Figure : Missing values in microarray data

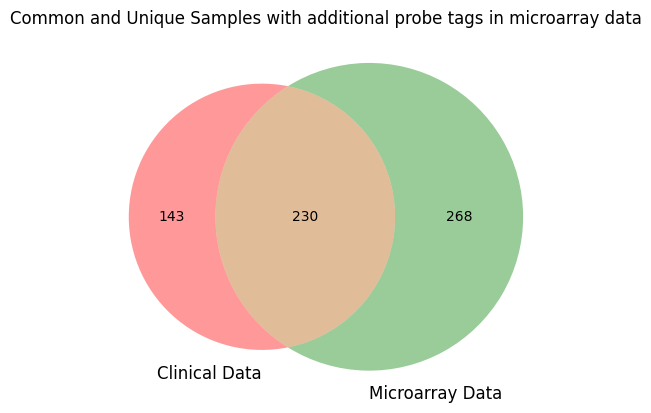
**Visualisation of common samples in clinical data and microarray data**

Figure :Venn diagram showing common samples among clinical and microarray data

Since the clinical data contains only 373 samples while the microarray data has 498 samples, visualizing the common samples using a Venn diagram can help us understand the overlap and differences between these datasets. Interestingly, only 230 samples are common whereas, 143 samples are unique to clinical data and 268 samples unique microarray data. The reason is that, as explained in the previous section, some of the samples (probe names) in the microarray data have additional tags which need a removal in data preprocessing.

**2. Data preprocessing**

**2.1. One-hot encoding the categorical features**

Categorical features are variables that represent categories or labels, such as colors, types of animals, or cities. Machine learning algorithms require numerical data to perform calculations effectively, so converting these categorical features into numerical form is essential. One-hot encoding is a crucial technique used when dealing with categorical features in machine learning. One-hot encoding assigns a unique binary variable (0 or 1) to each category within a feature. For example, if we have a "Gender" feature with categories like "Female" and "Male" one-hot encoding transforms it into separate binary columns (e.g., "Is\_Female" "Is\_Male"). This process ensures that the algorithm understands the distinctions between different categories without implying any ordinal relationship. It's a fundamental step in preprocessing data for machine learning, enabling models to work with a wider range of feature types and produce more accurate predictions. In Figure 33, clinical features before encoding can be seen and after one-hot encoding application for each categorical clinical feature, updated clinical data can be seen in Figure 34. During hyperparameter tuning one-hot encoding for multiclass INSS.Stage clinical endpoint was also applied.

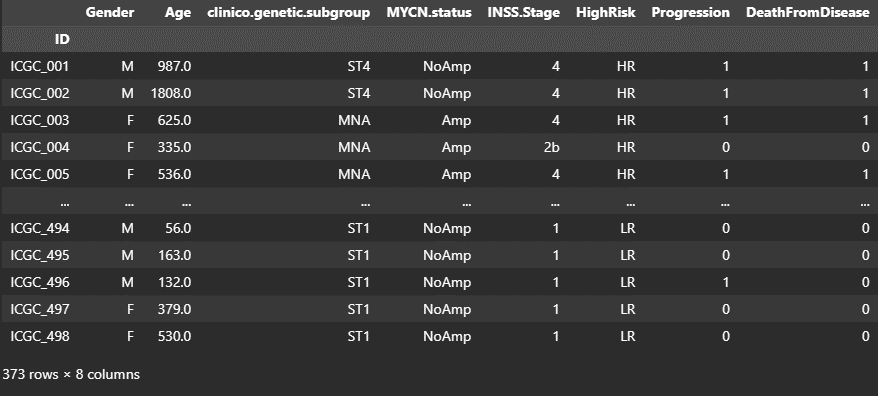
****

Figure : General overview of clinical data before one-hot encoding

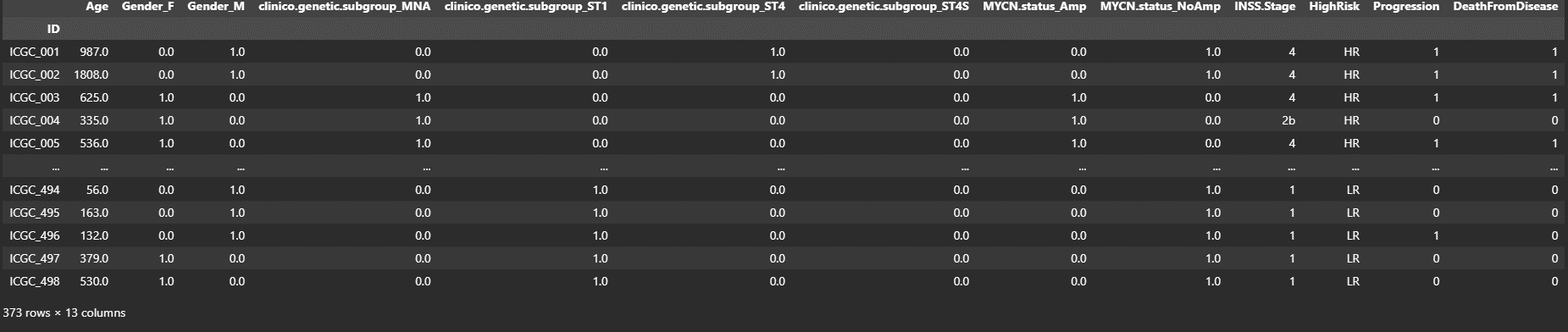


Figure : General overview of clinical data after one-hot encoding to categorical features

**2.2. Merging two data frames**

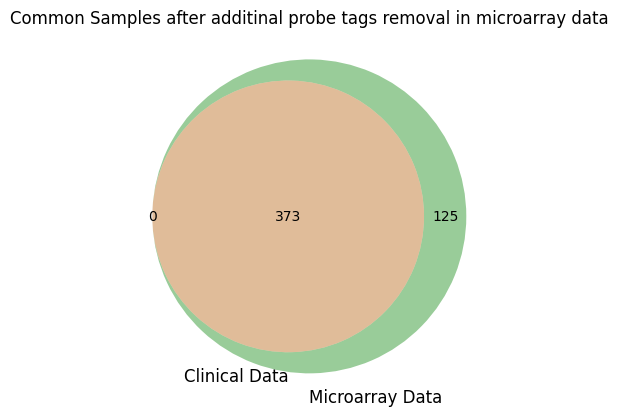


Figure :Venn diagram showing common samples after additional probe tags removal from microarray data.

After removal of additional tags from probe names in microarray dataset, the clinical and microarray data were merged based on their common samples (Figure35).

**2.3. Removing gene columns with complete missing values**

During, analyzing of missing values in microarray data, various missing values detected in microarray data. But interestingly, after more careful examination it was seen that some genes don't have expression values for any microarray probes (Figure 36).

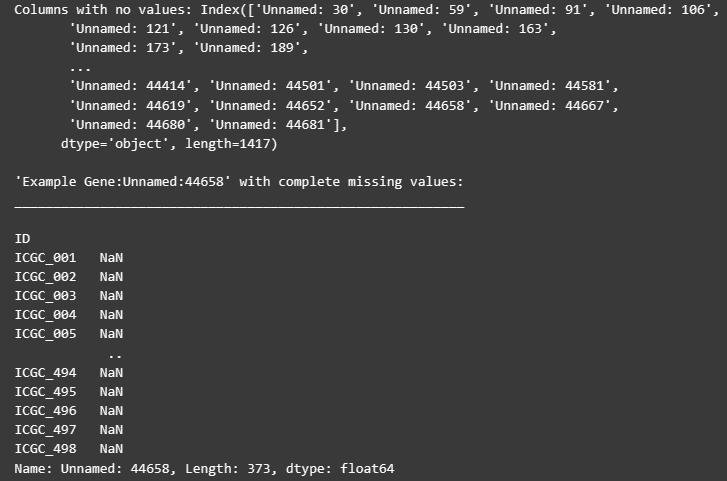


Figure : Gene columns without expression values and one example of them (Unnamed:44658)

Because complete missingness in the entire column, these values cannot be imputed and therefore these columns were removed from the dataset. After removal 43291 genes were left in the dataset (Figure 37).

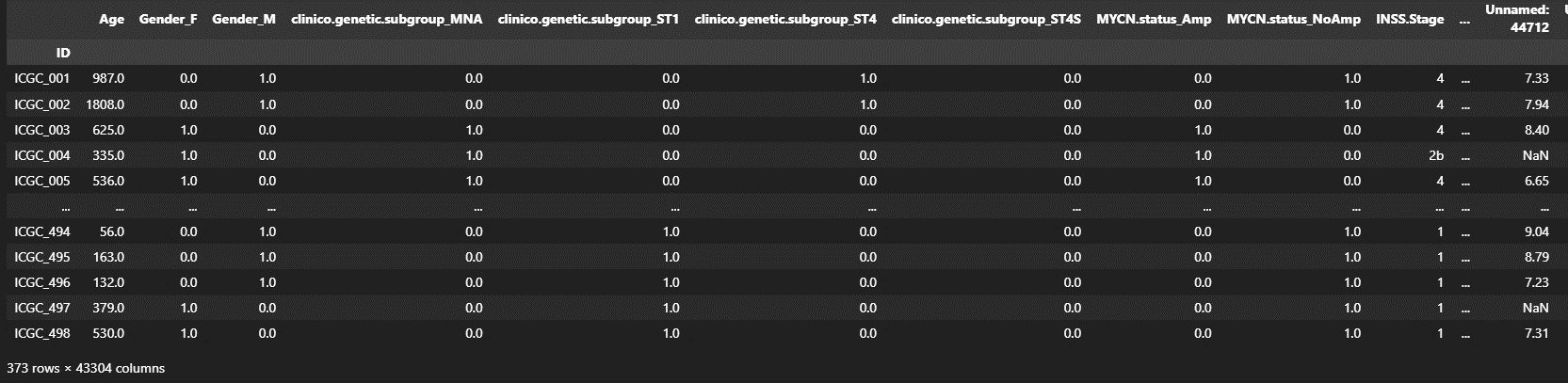


Figure : Merged dataset after removal of gene columns having no gene values. (43304 feature column = clinical features + clinical endpoints + 43291 genes)

**2.4. Splitting the data as train and test**

As next, feature matrix, which contains all the features and clinical end point matrix, containing all clinical endpoints are assigned to new data frames as X and y. Then X and y are randomly split to train and test data such that test data contains 20% of the main dataset.



Figure : Dimension of Feature matrices after splitting

**2.5. Data imputation**

While there is no missing value in clinical data (features and clinical endpoints), there are various missing gene expression values in microarray data (Figure39).

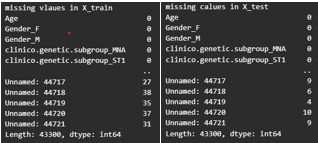


Figure :Missing values in gene features indicating in both X\_train and X\_test

In order to impute the microarray data missing values, microarray data is subset to new dataframe and then imputed by using the strategy as filled the values with “median” column wise both for X\_train and X\_test datasets separately. Afterwards, the subset data frames were merged to clinical data again (Figure 40).

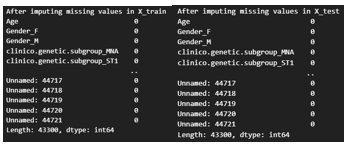


Figure ::After data imputation, there is no missing values in gene features indicating in both X\_train and X\_test

**2.6. Data normalisation**

One-hot encoding is specifically designed for categorical data and, as such, there's no need to apply data normalisation to one-hot encoded features (Stackexchange). However, it's crucial to normalize the rest of your numerical data. Normalization ensures that numerical features are on a consistent scale, preventing some variables from dominating the model due to their larger values. I, therefore, performed normalization as standard scaling, only on the subsets of numerical categories within X\_train and X\_test. I also examined the Age distribution alongside one randomly selected gene. As I anticipated, the Age distribution displayed skewness after normalization due to the outliers (Figure 41).

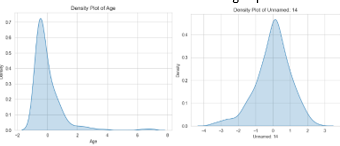


Figure : Age and Unnamed:14 distributions after normalisation

Finally, the normalized features were reintroduced into X\_train and X\_test (Figure 42).

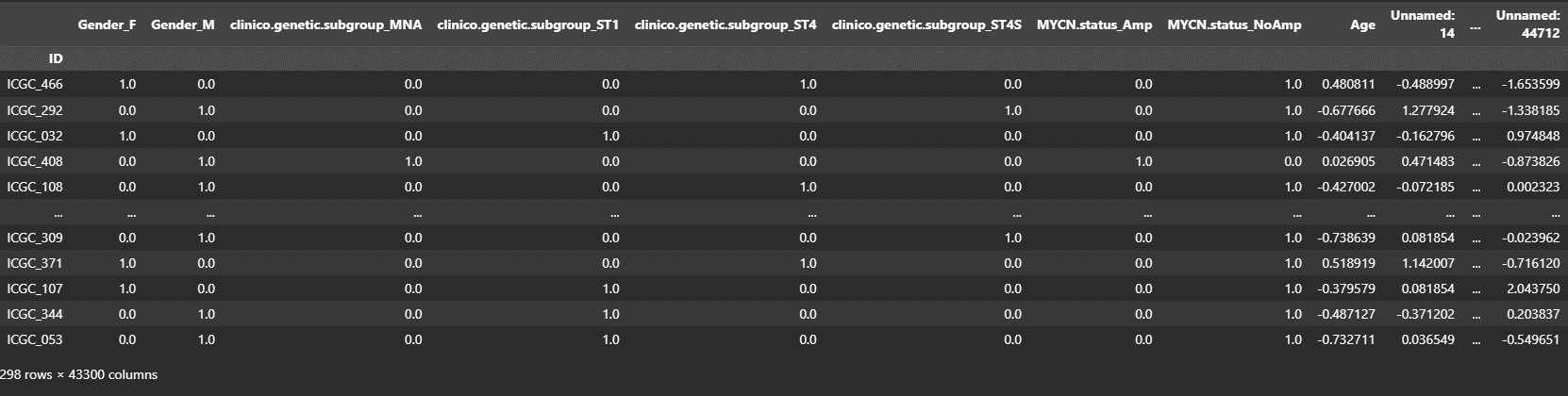


Figure : X\_train feature matrix after normalisation

**3. Dimensional Reduction to analyse clinical endpoint separation**

In order to understand the clusters and patterns in high dimensional X\_train dataset (298x43400) (Figure 42), a dimensional reduction technique: PCA (Principle Component Analysis) and Kmeans clustering were performed. PCA is a dimensionality reduction technique used to project the high dimensional data in new dimensions called as principe components for simplifying data interpretation and visualization. It helps capture the most important information in the data while minimizing information loss. In addition to PCA, K-means is an unsupervised clustering algorithm, aimed at grouping data points into distinct clusters based on their similarities. When PCA and K-means are combined, it helps to uncover underlying patterns in high-dimensional data.

**3.1. INSS.Stage**

**3.1. 1. PCA for INSS.Stage**

After performing PCA for X\_train and label the data points for INSS.Stage, it is was seen that there was no clear separation among subclasses in the lower-dimensional representation (Figure 43). One reason is that microarray data by nature may not be linearly separable thus PCA is not suitable. However, it might also suggest that the data might not exhibit distinct clusters or patterns that can be effectively separated using PCA alone. After inspecting the subclasses in INSS.Stage it was seen that not all stages do not have representative samples (see the EDA section INSS.Stage and Figure 43). This might explain why the data, projected on lower dimension didn’t separate in clusters. The suggestions to overcome this issue is following: Collecting more data for different INSS.Stage subgroups (there is only one sample for subgroups in stage2, 2a and 2b and 9 samples for Stage3) and if this is not possible to combining subgroups (For example stage2, Stage2a and Stage2b) or perform oversampling for underpresented samples. However, because there is only one sample representation for stage2, 2a and 2b, it is not recommended to perform oversampling since this is not a representative number of samples for the particular stage.

|  |  |
| --- | --- |
| **Visualization different combinations of principal components for INSS.Stage** | |
|  |  |
|  |  |
|  |  |

Figure : Visualisation of different combinations of principle components for INSS.Stage

**3.1.2 Kmeans clustering for INSS.Stage**

In order to understand to visualise natural groupings or patterns in the X\_train feature matrix, I performed Kmeans by randomly selecting two genes from microarray data (Figure 44 left side). I also compared the predicted labels by K-means with the original labels of INSS.Stage clinical endpoint. From 96 Stage 1 class label (1), Kmeans predicted 79 of them correctly. From 1 Stage2 class label (2), Kmeans predicted 0 of them correctly. From 1 Stage2a class label (2a), Kmeans predicted 0 of them correctly. From 1 Stage2b class label (2b), Kmeans predicted 1 of them correctly. From 9 Stage3 class label (3), Kmeans predicted 6 of them correctly. From 151 Stage4 class label (4), Kmeans predicted 61 of them correctly. And lastly, From 39 Stage4S class label (4S), Kmeans predicted 28 of them correctly. These results suggest that K-means didn’t perform very good by using these two genes to INSS.Stages. In addition to this, I also overlayed the K-means labels and original class labels on the PCA result (Figure 44 right side). Although I observed a nice separation in lower dimension, overlay of K-means labels with the original class labels showed several mismatches. These results tell that K-means performance was not well indicating that these two features are not strong indicators of the outcome or that more features are needed for accurate predictions.

|  |  |
| --- | --- |
| **KMeans Clustering for INSS.Stage by using randomly picked two genes from microarray data** | |
|  |  |

Figure : KMeans clustering for INSS.Stage by using two randomly selected genes

**3.2. HighRisk**

**3.2.1** **PCA for HighRisk**

PCA for X\_train and labeling the data points with HighRisk clinical endpoint provided more promising results compared to INSS.Stage (Figure 45. Interestingly, although most of the variance of the data was projected on PC1 and PC2, more clear separation observed in PC1-PC3 and PC2-PC3. This can be explaining the following reasons. Although the variance along PC1 and PC2 may not necessarily align with the variations that are most critical for distinguishing between classes (HighRisk and LowRisk). When the PC3 component is integrated, more clear separation was seen indicating that PC3 captures more relevant part of the data for the class separation. Another reason would be that the nature of the relationship between the features in the dataset and the target variable (HighRisk) may not be linear or easily captured by only the top two principal components. Whereas PC3 could capture nonlinear relationships or other critical information. Furthermore, unlike INSS.Stage, subclasses (HR(highrisk) and LR(lowrisk)) are more equally representative in HighRisk clinical endpoint indicating no class imbalance. This means that both subclasses have a better chance of being represented in the principal components, potentially leading to improved classification results.

|  |  |
| --- | --- |
| **Visualization different combinations of principal components for HighRisk** | |
|  |  |
|  |  |

Figure :Visualisation of different combinations of principle components for HighRisk

**3.2.2** **Kmeans clustering for HighRisk**

To reveal inherent data patterns and natural groupings, I employed K-means clustering, selecting two genes randomly from microarray data (as depicted in Figure46 on the left). Furthermore, I contrasted K-means' predicted labels with the original labels of the HighRisk clinical endpoint. From 161 Lowrisk class labels (“LR”), Kmeans predicted 78 of them correctly. And From 137 Highrisk class labels (“HR”), Kmeans 116 of them correctly. Here we can say that Kmeans had a very good performance predicting HighRisk patients compared to LowRisk patients. This also suggests that the microarray data is reduced to just these two genes, may have more distinctive patterns for "HighRisk" patients, making it easier for K-means to separate them into a distinct cluster. On the other hand, the "LowRisk" patients might exhibit more variation in gene expression within their group, making it less straightforward for K-means to separate them. In addition to this, I also overlayed the K-means labels and original class labels on the PCA result (Figure46 on the right). Although we observed a nice separation in lower dimension, K-means label overlay with the original class labels showed several mismatches however the overall result is much better than INSS.Stage. These results tell that K-means performance overall didn’t perform very well indicating that these two features are not strong indicators of the outcome or that more features are needed for accurate predictions.

|  |  |
| --- | --- |
| **KMeans Clustering for HighRisk by using randomly picked two genes from microarray data** | |
|  |  |

Figure :KMeans clustering for HighRIsk by using two randomly selected genes

**3.3. Progression**

**3.3.1. PCA for Progression**

Similar to PCA result of HighRisk, PCA for X\_train and and labeling the data points with Progression clinical endpoint provided more promising results compared to INSS.Stage (Figure 47). And again, similar to HighRisk, although most of the variance of the data was projected on PC1 and PC2, more clear separation observed in PC1-PC3. As reasons explained above, it seems like PC3 captures more critical data part for distinguishing the Progression class. Furthermore, unlike INSS.Stage, subclasses (0 and 1(progression of cancer)) are slightly better representative in Progression clinical endpoint indicating class imbalance may not be too worrying. This means that both subclasses will have a chance of being represented in the principal components, potentially can help classification results.

|  |  |
| --- | --- |
| **Visualization different combinations of principal components for Progression** | |
|  |  |
|  |  |

Figure :Visualisation of different combinations of principle components for Progression

**3.3.2. Kmeans clustering for Progression**

To unveil inherent data patterns and natural groupings, I conducted K-means clustering by randomly selecting two genes from microarray data (Figure48 left). Additionally, I juxtaposed K-means' predicted labels with the original labels of the Progression clinical endpoint. From 181 no progression class labels (“0”), Kmeans predicted 62 of them correctly. And from 117 progression class labels (“1”), Kmeans 56 of them correctly. In addition to this, I also overlayed the K-means labels and original class labels on the PCA result (Figure48 right). Although we observe a nice separation in lower dimension, K-means label overlay with the original class labels showed several mismatches. These results tell that K-means performance was not well indicating that these two features are not strong indicators of the outcome or that more features are needed for accurate predictions.

|  |  |
| --- | --- |
| **KMeans Clustering for Progression by using randomly picked two genes from microarray data** | |
|  |  |

Figure :KMeans clustering for Progression by using two randomly selected genes

**3.4. DeathFromDisease**

**3.4.1. PCA for DeathFromDisease**

Unlike Progression and HighRisk, PCA did not perform well on DeathFromDisease as the classes (0 and 1) were not separated clearly (Figure 49). One reason would be is that class representations in death and no death are different in the dataset: Neuroblastoma patients who do not die are much higher than the neuroblastoma patients who die (Figure 49). This imbalance class distribution can also cause worse classifier performance.

|  |  |
| --- | --- |
| **Visualization different combinations of principal components for DeathFromDisease** | |
|  |  |
|  |  |

Figure :Visualisation of different combinations of principle components for DeathFromDisease

**3.4.2. Kmeans clustering for DeathFromDisease**

To explore inherent data patterns and groupings, I conducted K-means clustering after randomly selecting two genes from microarray data (Figure 50 left). Additionally, I cross-referenced the labels generated by K-means with the original labels from the DeathFromDisease clinical endpoint. From 220 no deathclass labels (“0”), Kmeans predicted 80 of them correctly. And From 78 death class labels (“1”), Kmeans 35 of them correctly. In addition to this, I also overlay the K-means labels and original class labels on the PCA result (Figure 50 right). Although we observe a nice separation in lower dimension, K-means label overlay with the original class labels showed several mismatches. These results tell that K-means performance was not well indicating that these two features are not strong indicators of the outcome or that more features are needed for accurate predictions.

|  |  |
| --- | --- |
| **KMeans Clustering for DeathFromDisease by using randomly picked two genes from microarray data** | |
|  |  |

Figure :KMeans clustering for DeathFromDisease by using two randomly selected genes

**4. Feature importance and Feature selection**

Since there are 43300 feature columns in X matrix, I employed both feature importance and a feature selection methodology to X\_train feature matrix identify a subset of informative features for classification to predict clinical endpoints.

**Feature importance:**

I conducted an in-depth analysis of feature importance aimed to identify the top 100 features for a classification task, aiming to unveil the critical factors that influence the model's decision-making process. To achieve this, I employed a Random Forest Classifier, a powerful ensemble learning technique, combined with a Stratified K-Fold cross-validation strategy with five splits ensuring that each fold preserved the original class distribution. For each cross-validation split, the Random Forest Classifier was trained and feature importances were extracted. By calculating the mean feature importances and sorting them in descending order, I obtained a list of critical features common among folds.

**Feature selection:**

In order to select features to improve model performance and interpretability, I performed feature selection method. For this, I again utilized a Stratified K-Fold cross-validation approach with five splits to ensure robustness and unbiased evaluation. To select the relevant features, I used SelectKBest and SelectPercentile, which are feature selection methods that evaluate feature importance based on statistical metrics. The top 300 common features were chosen as they consistently demonstrated strong discriminative power across different folds. I afterwards trained a Random Forest Classifier on each split using the selected features and computed the different metrics (f1 and accuracy) to evaluate the classification performance.

**4. 1. INSS.Stage**

Because the INSS.Stage is related to the clinico.genetic.subgroup feature, all one-hot encoded subgroups of this feature were removed from the feature matrix for feature importance and feature selection of INSS.Stage. In addition, INSS.Stage as y\_train is also encoded by one-hot-encoding method due to its multiclass classification nature.

**4. 1.1. Feature importance for INSS.Stage**

Interestingly only 8 features were common among the most 100 importance features in each fold (Figure 51). Furthermore, with the exception of the 'KCNK18' gene, all other features consistently selected in the top 100 were determined by the feature selection process, as evidenced in the feature selection section below (Figure 52) . This observation underscores the robust performance achieved through both methods—feature selection and the Random Forest Classifier, collectively contributing to the model's predictive power.

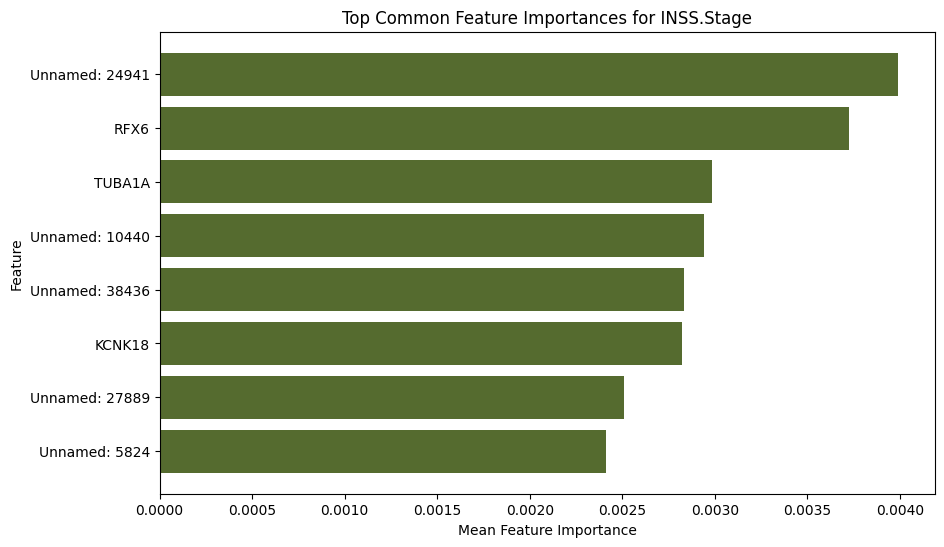
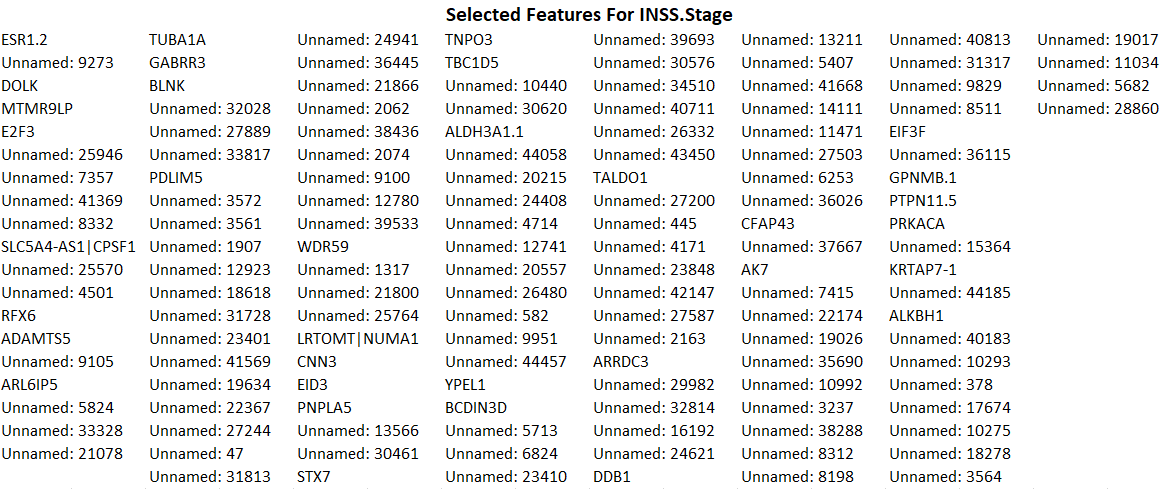


Figure Top Common Feature importances for INSS.Stage

**4. 1.2. Feature selection for INSS.Stage**

The results indicated an average accuracy of 0.73 and an average F1 score of 0.68. At the end, a total of 143 features were selected as informative for the classification task, (Figure 52). Interestingly, it's worth noting that among the selected features, none of them were clinical features, indicating that the model relied primarily on the microarray data. Lastly, Random Forest model evaluation scores (accuracy and f1 score) indicate a relatively fine model performance with these features. Although 'KCNK18', the gene was selected in top feature importance, it was not selected in the feature selection method. However, this feature can also have a potential power for increase model performance on prediction to INSS.Stages. Therefore, it can be integrated to feature matrix and model performance can be evaluated together with this feature.

Figure :Selected features for INSS.Stage



**4. 2. HighRisk**

**4. 2. 1. Feature importance for HighRisk**

14 features were seen to be common among the most 100 importance features in each fold (Figure 53). Although there are more common features among folds, the features that are found to be important: Unnamed: 31382, Unnamed: 11592, Unnamed: 6253, Unnamed: 15515, Unnamed: 2808, Unnamed: 38423 were not selected in feature selection method (Figure 54). It's possible that the feature selection methods (SelectPercentile, KBest) may not have recognized the importance of these specific features, while the Random Forest Classifier found them to be valuable.

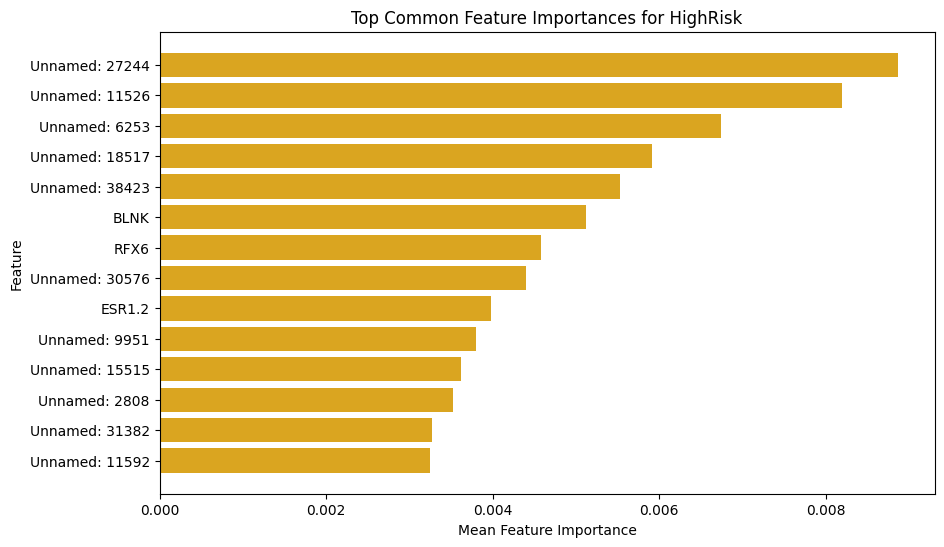
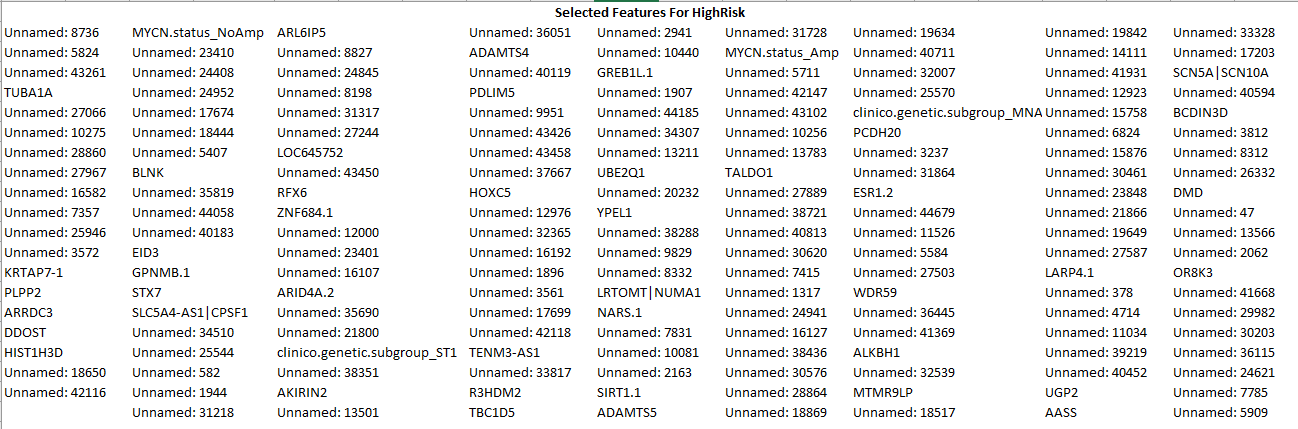
****

Figure :Top Common Feature importances for HighRisk

**4. 2. 2. Feature selection for HighRisk**

A total of 179 features were selected as informative for the classification task (Figure 54). As expected (explained in EDA section), MYCN.status (NoAmp and Amp) and clinico.genetic.subgroup (ST1 and MNA) were also inside in the selected features. Although many genes thought to be important didn’t selected during feature selection. Lastly, Random Forest model evaluation scores (accuracy: 0.92 and f1 score:0.92) indicate a very good model performance with these selected features.

Figure :Selected features for HighRisk



**4. 3. Progression**

**4. 3. 1. Feature importance for Progression**

11 features were seen to be common among the most 100 importance features in each fold (Figure 55 ). Among these features, only Unnamed: 18517, Unnamed: 7285, Unnamed: 25494, Unnamed: 4453 and, Unnamed: 30866 were also selected by feature selection methods (Figure 56). It's possible that the feature selection methods (SelectPercentile, KBest) may not have recognized the importance of these specific features, while the Random Forest Classifier found them to be valuable.

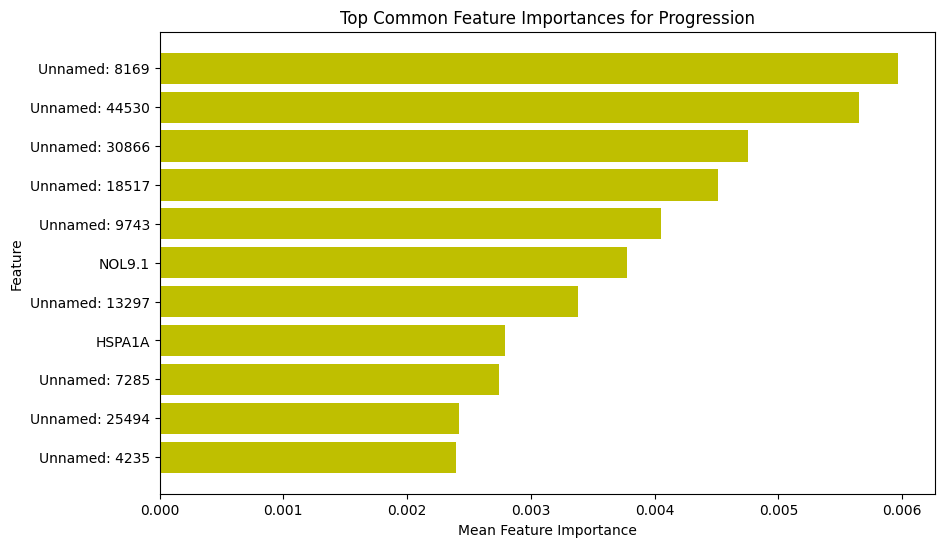
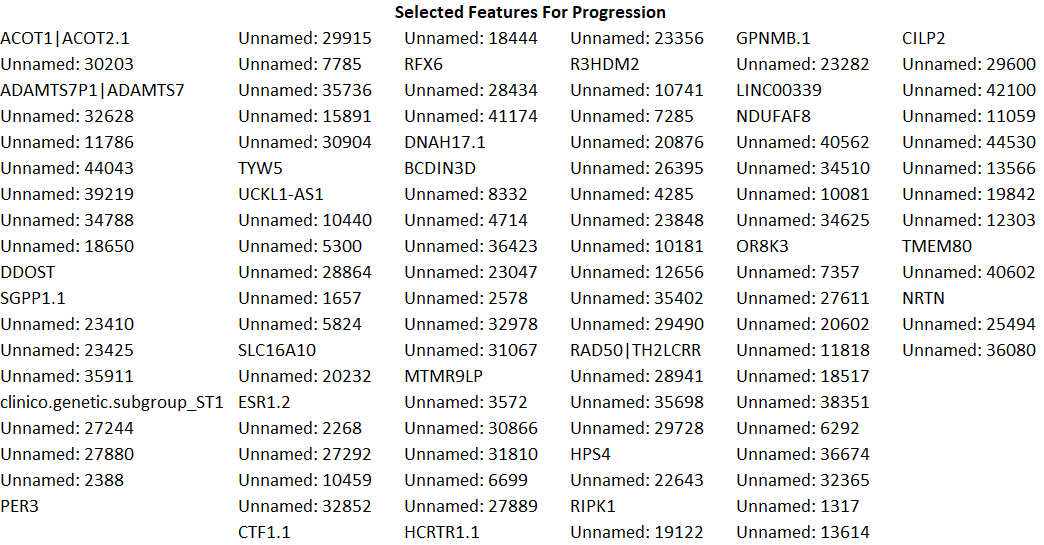
****

Figure ::Top Common Feature importances for Progression

**4. 3. 2. Feature selection for Progression**

A total of 112 features were selected as informative for the classification task (Figure 56). As expected clinico.genetic.subgroup was also selected within these selected features (reasoning explained in EDA section). Finally, model performance evaluation of a Random Forest Classifier showed the accuracy and f1 score as 0.75 for both with a relatively fine model performance. To improve the model performance, these features: Unnamed: 9743, Unnamed: 13297, NOL9.1, HSPA1A, Unnamed: 4235 and, Unnamed: 8169 which were found as important features can be integrated to feature matrix and model performance can be evaluated with these features.

Figure :Selected features for Progression



**4.4. DeathFromDisease**

**4.4. 1. Feature importance for DeathFromDisease**

11 features were seen to be common among the most 100 importance features in each fold (Figure 57). Among these features, only Unnamed:7285, Unnamed: 8169, Unnamed: 28941, Unnamed: 29600, Unnamed:43240 were also selected by feature selection methods (can be Figure 58). It's possible that the feature selection methods (SelectPercentile, KBest) may not have recognized the importance of these specific features, while the Random Forest Classifier found them to be valuable.

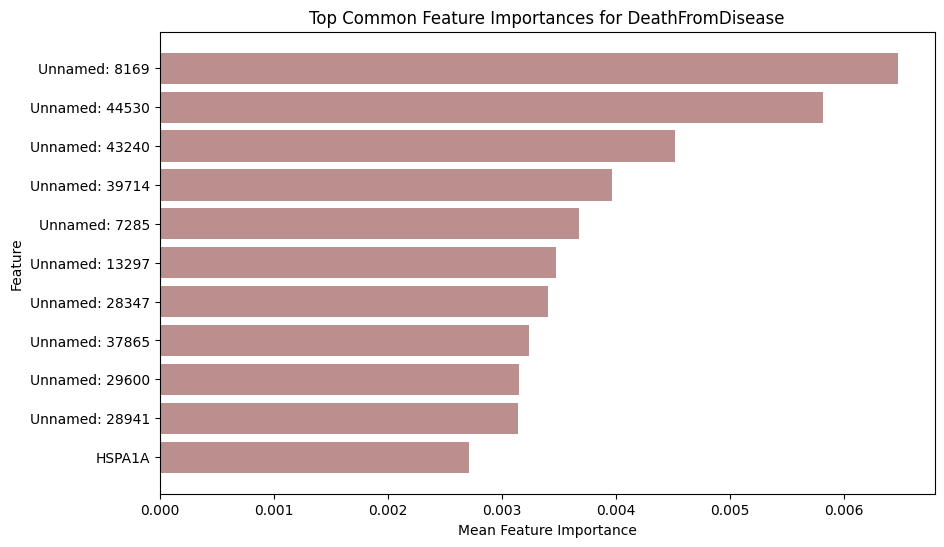
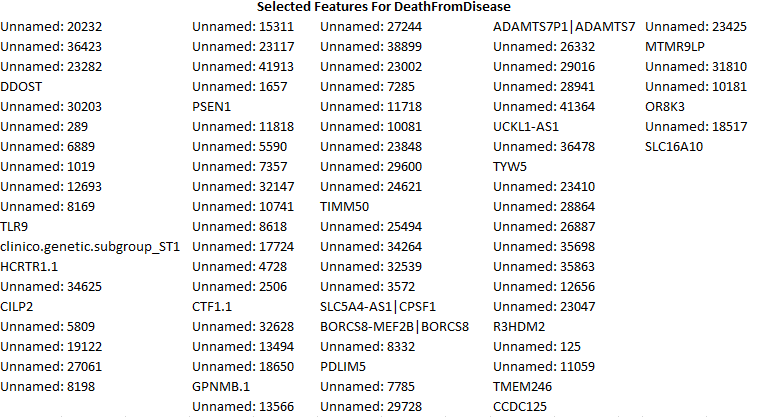
****

Figure :Top Common Feature importances for DeathFromDisease

**4.4. 2. Feature selection for DeathFromDisease**

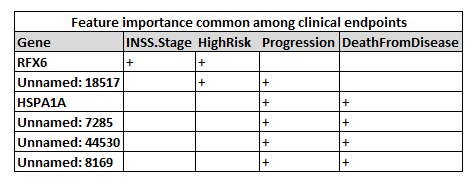
Interestingly only a total of 89 features were selected as informative for the classification task (Figure 58). As expected clinico.genetic.subgroup (ST1) was also selected within these selected features (explained in EDA section). Finally, model performance evaluation of a Random Forest Classifier showed the accuracy and f1 score as 0.79 and 0.78 respectively.To improve the model performance, these features: Unnamed:39714, Unnamed:13297, HSPA1A, Unnamed:44530, Unnamed:28347, Unnamed:37865 which were found as important features can be integrated to feature matrix and model performance can be evaluated with these features.

Figure :Selected features for DeathFromDisease



In summary, the evaluation of feature importance and feature selection methods revealed a substantial overlap in identified features. Additionally, my analysis identified specific features that are of significant importance in predicting clinical endpoints (Table1). Notably, these genes exhibit a common presence among patients classified as HighRisk, Progression, and DeathFromDisease, suggesting a potential association between these genes and disease severity.

Table1: Common Feature importances among clinical endpoints



Moreover, my feature selection analysis consistently emphasized the relevance of clinico.genetic.subgroup across various clinical endpoints, establishing its pivotal role in model training. In the case of HighRisk prediction, both clinico.genetic.subgroup and MYCN.status emerged as top-ranking features, highlighting the significance of clinical attributes in modeling.

**5. Hyperparameter tuning and stratified cross-validation**

After hyperparameter tuning, I conducted stratified cross-validation by using the X\_train feature matrix only. This practice allows me to evaluate the model's generalization performance within the training dataset without risking data leakage from the reserved test data. In addition, CV provides a valuable assessment of how well the model performs across different subsets of the training data and seeing the anomalies on the data which offering a reasonable estimate of its potential performance on unseen data. Futhermore, comparing the cross-validation results to the initial single random split helps me estimate the consistency of the model's performance and builds confidence in the reliability of that single split.

As my initial model selections, I opted for Random Forest, Radial Support Vector Machine, and Logistic Regression. And the Radial Support Vector Machine (SVM) was chosen due to its non-linear capabilities, making it a valuable addition to the model selection for this complex dataset. The algorithms are well-known for their effectiveness in handling wide datasets and often practiced the clinical data (Huang et al., 2018, Qian et al., 2021, Acharjee et al., 2020, Dai et al., 2018, Khairunnahar et al., 2019, Zangmo and Tiensuwan, 2018).

Hyperparameter tuning with Grid Search for both Random Forest and Logistic Regression and with Optuna (Akiba et al., 2019) for Radial Support Vector Machine were performed with 5 fold cross validation with predefined hyperparameter ranges. This was followed by a 3-fold cross-validation process on both a subset of X\_train data obtained from selected features and the complete X\_train dataset. The results are presented below.

**5. 1. INSS.Stage**

Confusion matrices, model performance metrics and best hyperparameter tunings for both selected and all feature matrices for INSS.Stage can be seen Figure 59-61. In general, both for selected features and all features of cross-validation results with the best parameters show quite good results as seen with different performance metrics. However, INSS.Stage clinical endpoint contains imbalanced classes. Although majority classes 1 and 4 mostly corrected very well, minority classes 2,3 and 4S were predicted very poorly which can be seen in confusion matrices. Handling these imbalanced classes with oversampling methods can help to improve prediction performances of classifiers. Moreover, the stratified cross-validation folds showed that there is no anomaly in the dataset, such that each cross gave similar results. In addition to this, classifiers were also performed similarly. Lastly, it is important to mention that because INSS.Stages are multiclass classification problems, for calculation of metrics, micro-averaging was used. Micro-averaging is particularly advantageous in the presence of class imbalance. When certain classes significantly outnumber others, micro-averaged metrics offer a more accurate and balanced evaluation of the model's overall performance (Kaggle). This was the case for INSS.Stage (See the EDA section). In summary, the results show that following hyperparameter tuning and cross-validation, all three classifiers performed well for the majority classes but less for minority classes. Additionaly, there was no significant difference in performance among the classifiers.

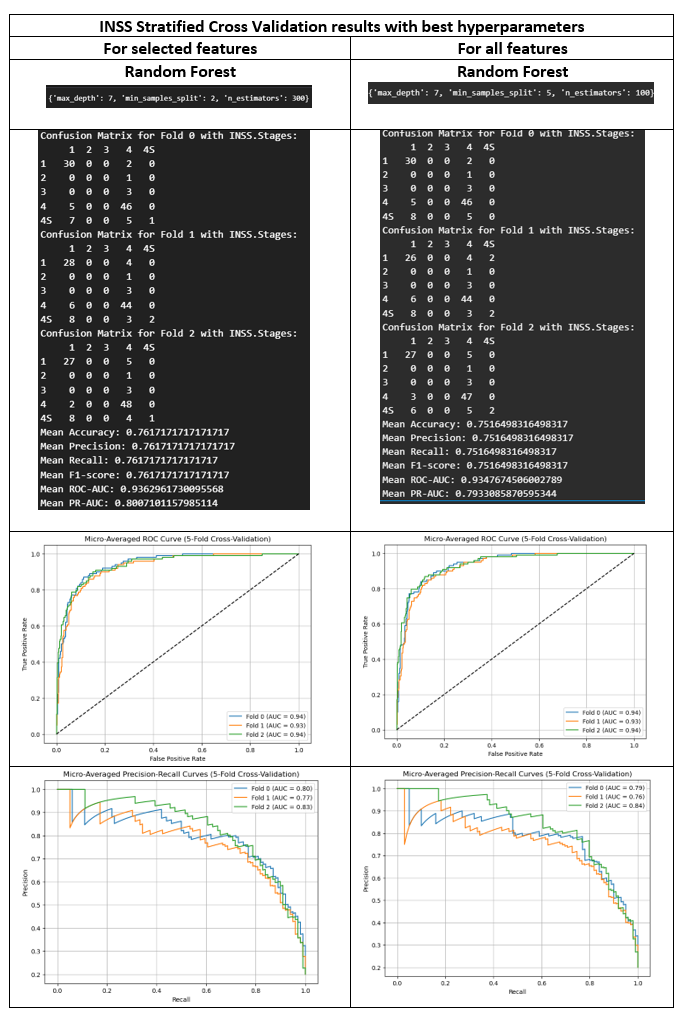
****

Figure :Best hyperporameters for Random Forest and Cross-Validation results for INSS.Stage - for selected and all features matrices

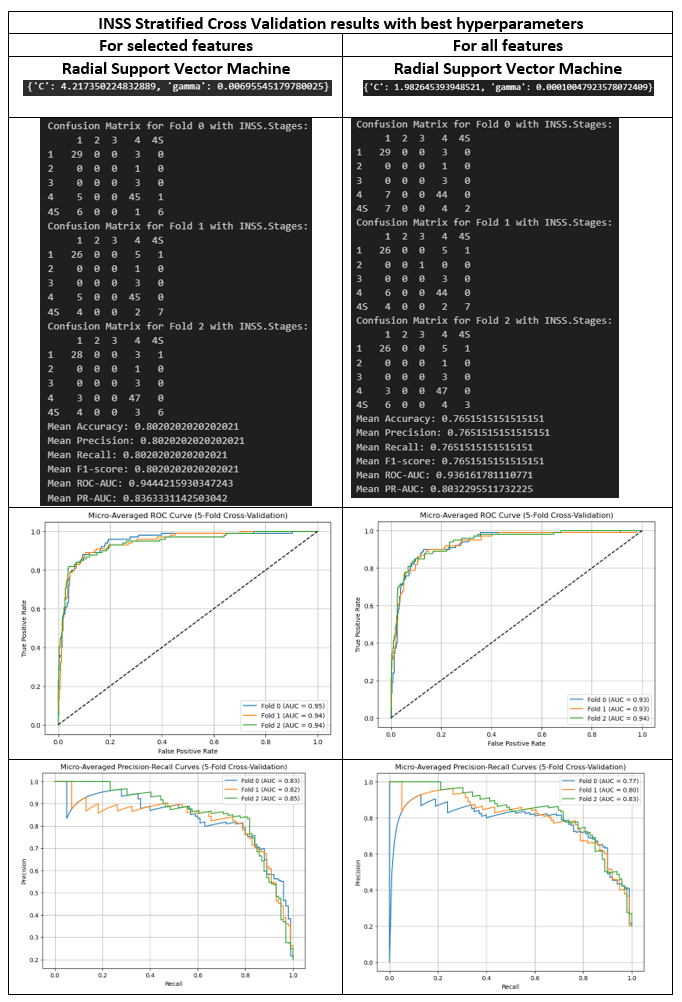
****

Figure :Best hyperporameters for SVM and Cross-Validation results for INSS.Stage- for selected and all features matrices

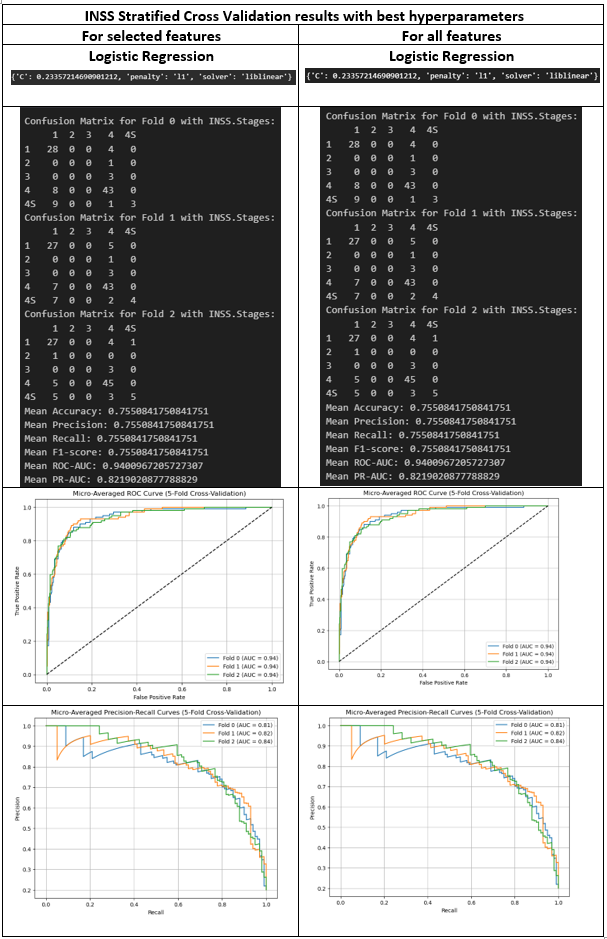
****

Figure :Best hyperporameters for Logistic Regression and Cross-Validation results for INSS.Stage - for selected and all features matrices

**5. 2. HighRisk**

Confusion matrices, model performance metrics and best hyperparameter tunings for both selected and all feature matrices for HighRisk can be seen in Figure 62-64. In general, both for selected features and all features of cross-validation results with the best parameters show outstanding results as seen with different performance metrics: confusion matrices, accuracy, precision, recall, and F1-score, ROC-AUC and PR-AUC (all around 90% and above). Furthermore, HighRisk clinical endpoint, unlike INSS.Stages do not contain imbalanced classes. This significantly affected the model performances. Both HR and LR classes predicted very well that can be seen in confusion matrices. This emphasizes the importance of balanced classes in clinical endpoints. Furthermore, similar to INSS.Stages, the stratified cross-validation folds showed that there is no anomaly in the dataset, such that each cross, less gave similar results. Lastly, in contrast to INSS.Stages, HighRisk was a binary classification problem (See the EDA section). Therefore, for model performance metrics, binary averaging was used. In summary, the results show that, cross-validation results with tuned hyperparameters of all three classifiers performed extremely well. It is also important to mention that there was no significant difference in performance among the classifiers, highlighting the robustness of these models in achieving accurate classification results with balanced classes in HighRisk.

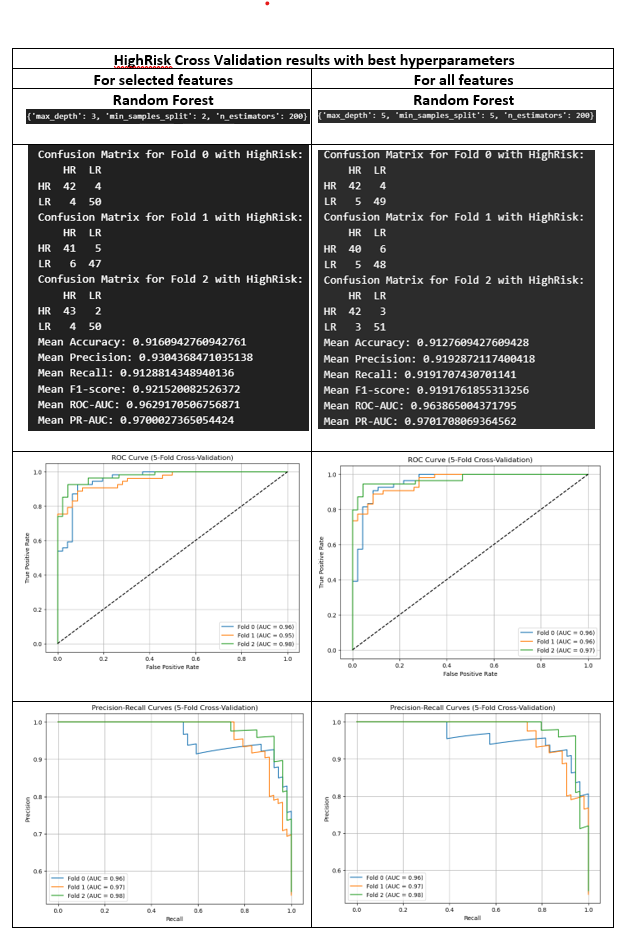
****

Figure :Best hyperporameters for Random Forest and Cross-Validation results for HighRisk- for selected and all features matrices

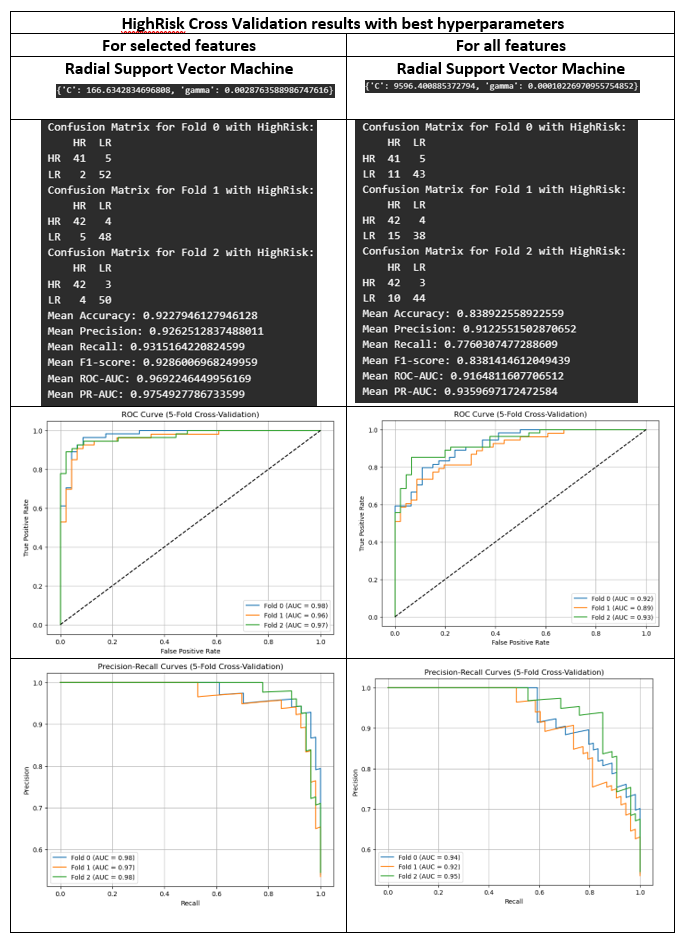


Figure ::Best hyperporameters for SVM and Cross-Validation results for HighRisk- for selected and all features matrices

****

Figure : Best hyperporameters for Logistic Regression and Cross-Validation results for HighRisk - for selected and all features matrices

**5. 3. Progression**

Confusion matrices, model performance metrics and best hyperparameter tunings for both selected and all feature matrices for Progression can be seen in Figure 65-67. In general, both for selected features and all features of cross-validation results with the best parameters show good results as seen with different performance metrics: confusion matrices, accuracy, precision, recall, and F1-score expect for all feature SVM cross validation with the best hyperparameters where performance metrics; precision, recall and f1 score was 0 and accuracy was 0.60 (60% accuracy). This also reflected the in ROC-AUC with 50% meaning that classifier correct prediction like random guessing. Like INSS.Stages, Progression clinical endpoint also contains imbalanced classes. Although majority class no progression (0) seen more accurately predicted, whereas minority class was predicted poorly which can be seen in confusion matrices. This is even significantly visible in SVM modelling for by using all feature matrix. This again shows the importance of handling imbalanced samples within the target variable to help to improve prediction performances of classifiers. In addition, there was no anomaly among different folds was seen during cross-validation. Lastly, Progression clinical endpoint like HighRisk is a binary class classification problem therefore for model performance metrics, binary averaging was used. In summary, the results show that cross-validation results with tuned hyperparameters of three classifiers perform fine except with the all-feature matrix combined with SVM which performed badly.

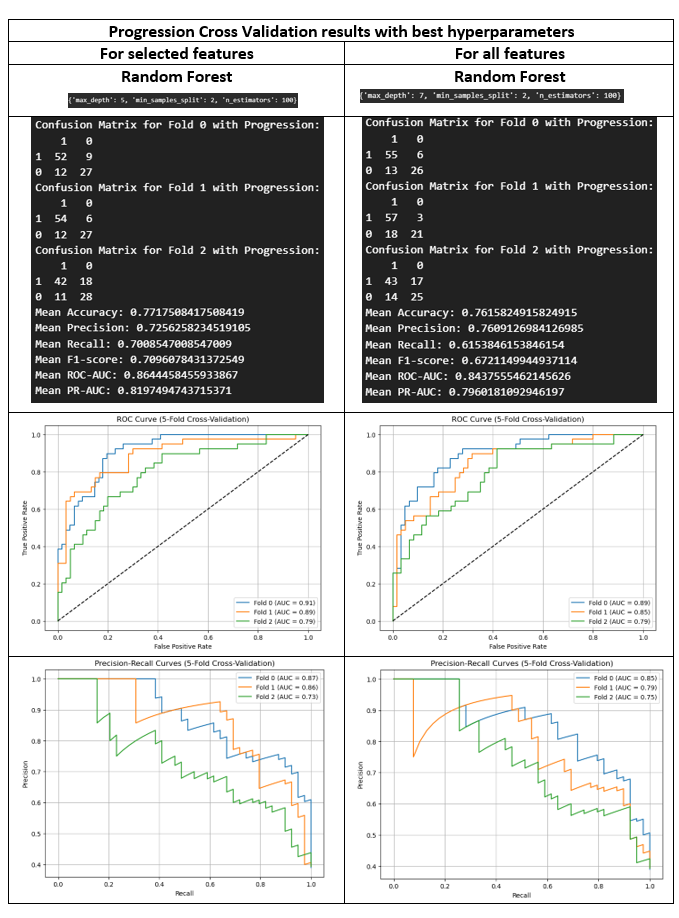
****

Figure : Best hyperporameters for Random Forest and Cross-Validation results for Progression - for selected and all features matrices

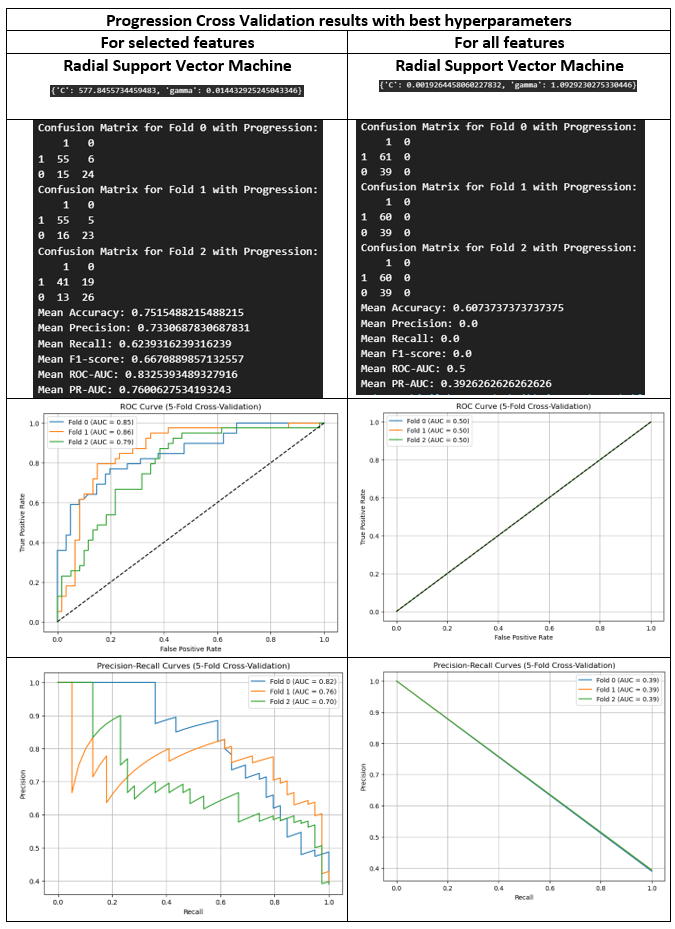
****

Figure : Best hyperporameters for SVM and Cross-Validation results for Progression - for selected and all features matrices

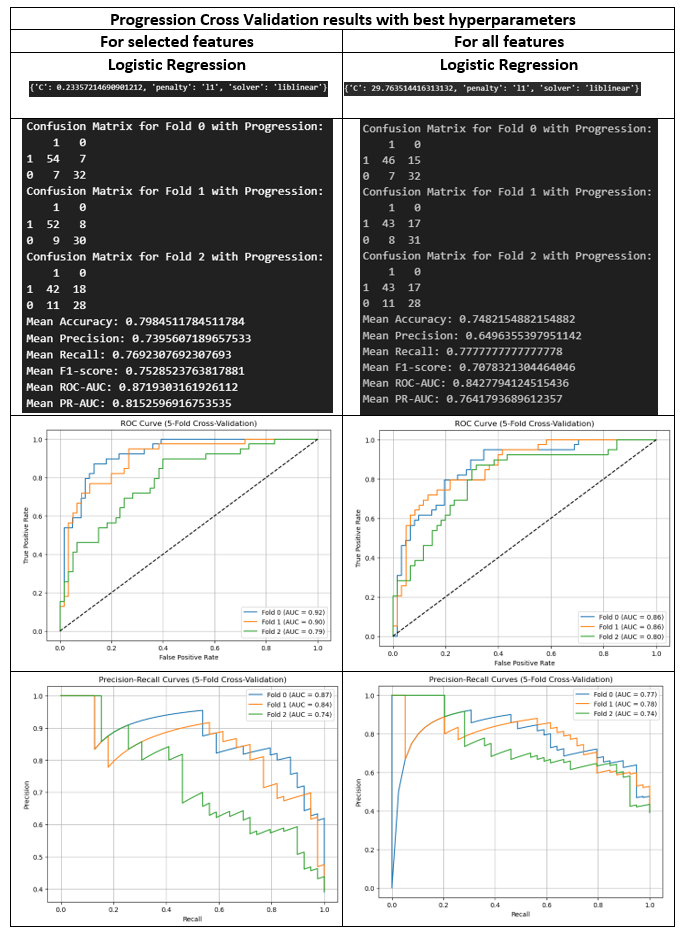
****

Figure 67: Best hyperporameters for Logistic Regression and Cross-Validation results for Progression - for selected and all features matrices

**5. 4. DeathFromDisease**

Confusion matrices, model performance metrics and best hyperparameter tunings for both selected and all feature matrices for DeathFromDisease can be seen in Figure 68-70. In general, similar to Progression, both for selected features and all features of cross-validation results with the best parameters show good results as seen with different performance metrics: confusion matrices, accuracy, precision, recall, and F1-score expect for all feature SVM cross validation with the best hyperparameters where performance metrics; precision, recall and f1 score was 0 and accuracy was 0.73 (73% accuracy). This also reflected the in ROC-AUC with 25% meaning that classifier correct prediction even worse than random guessing. Similarly, PR-AUC, a tradeoff precision and recall was only 0.18. This is because only no death from disease “0” subclass class was correctly predicted. This is because this clinical endpoint also significantly imbalanced and this affected the model performances. In this case, handling this imbalanced samples with oversampling methods can help to improve prediction performances of classifiers. Moreover, the stratified cross-validation folds showed that there is no anomaly in the dataset, such that each cross, less gave similar results. In addition to this, classifiers were also performed similarly. Lastly, Deathfromdisease like Progression and HighRisk clinical endpoints is a binary class classification problem therefore for model performance metrics, binary averaging was used. In summary, the results show that cross-validation results with tuned hyperparameters of three classifiers perform fine except with the all-feature matrix combined with SVM which performed badly.

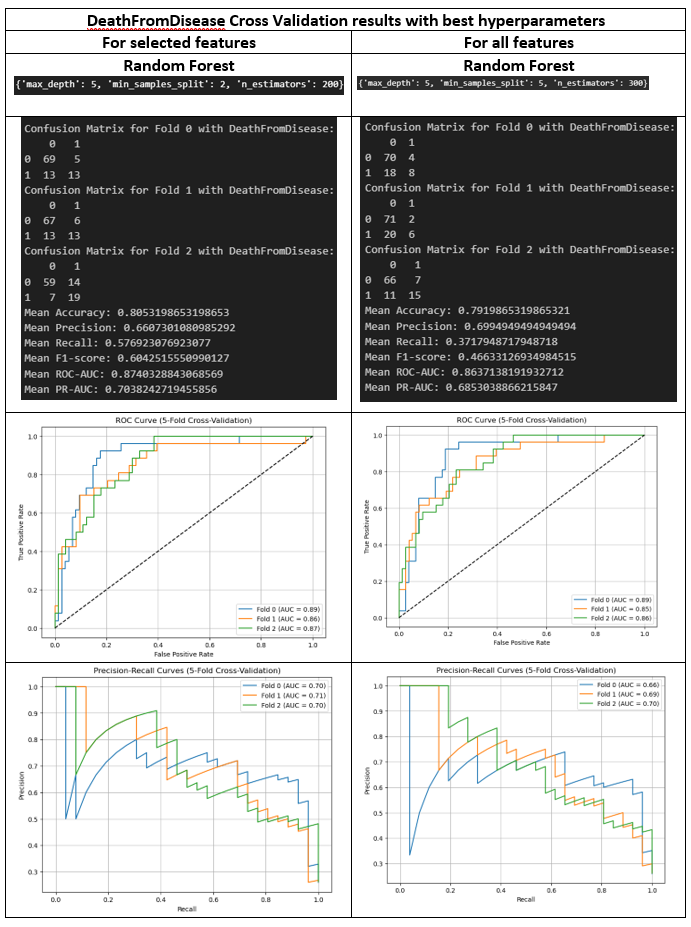
****

Figure 68: Best hyperporameters for Random Forest and Cross-Validation results for DeathFromDisease - for selected and all features matrices

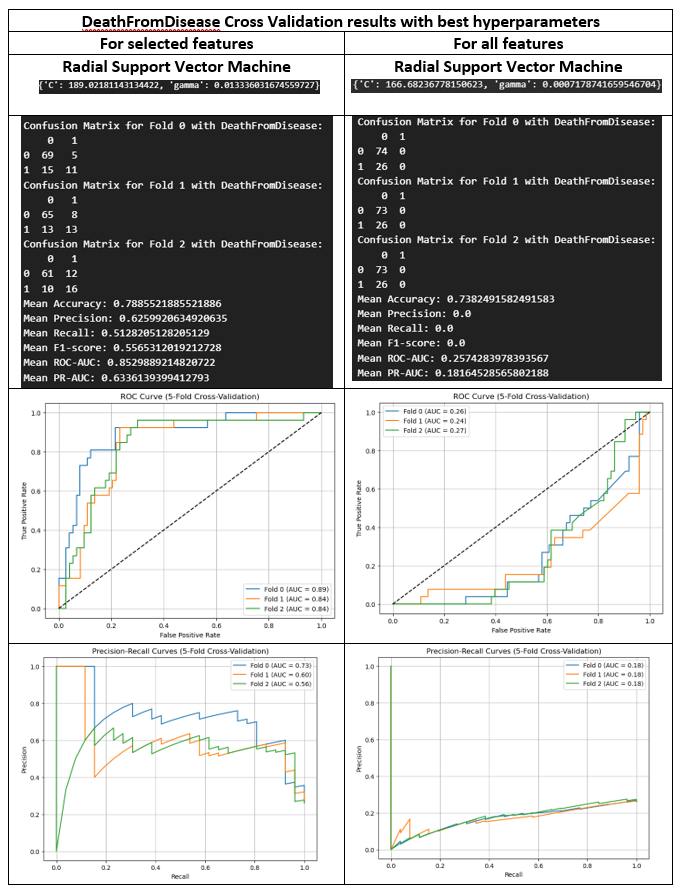
****

Figure 69: Best hyperporameters for SVM and Cross-Validation results for DeathFromDisease- for selected and all features matrices

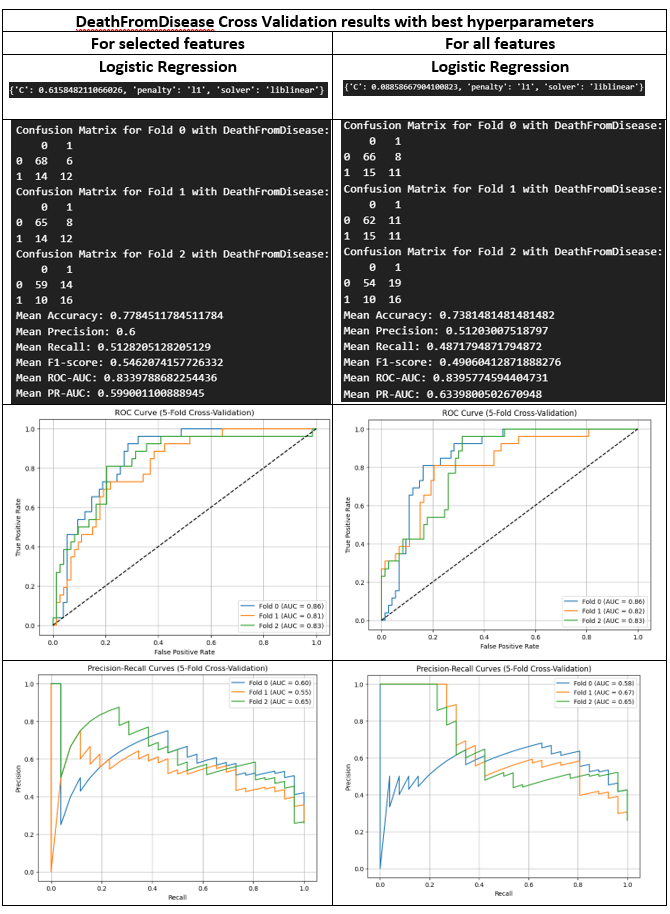
****

Figure 70: Best hyperporameters for Logistic Regression and Cross-Validation results for DeathFromDisease - for selected and all features matrices

**6. Model evaluation**

For model evaluation, all the models above were used to predict respected clinical endpoints with original unseen test data. The results of each clinical endpoints can be seen below.

**6.1 INSS**

INSS.Stage clinical endpoint has imbalanced subclass distributions (see the EDA section). Model evaluation for INSS.Stage reveals the presence of imbalanced subclass distributions, which can significantly influence the performance of machine learning models (Figure 71-73). In case of majority classes, the models consistently demonstrated robust performance (see the accuracy, ROC-AUC curve), reflecting their ability to accurately predict prevalent INSS stages. However, the situation differed when it came to minority classes (see the confusion matrices). The variation in model performance for these subclasses can be attributed to several factors. Firstly, limited sample size within minority classes poses a significant challenge for models to learn unique patterns for these subclasses which consequently reduces model accuracy. Another contributing factor is class imbalance. When the distribution of classes is highly skewed, models tend to become biased towards the majority class. This imbalance hampers the model's ability to differentiate between the underrepresented subclasses, resulting in suboptimal performance. Another important aspect is the relevance of features in the dataset. Some features may not be as informative for the minority classes, which can negatively impact the model's ability to make accurate predictions. In the feature importance part, I showed that some features are found the be important with clinical endpoints which are associated with disease severity. Maybe performing feature engineering to find features that are more representative of all subclasses (features that are less biased to clinical endpoint) can help to distinguish all subclasses. Nonetheless, it is evident that the number of representative samples for minority groups is limited. In cases where obtaining more data is not feasible, employing imbalanced sampling techniques like SMOTE can be beneficial. SMOTE can augment the sample size for imbalanced subgroups, facilitating the model's ability to capture patterns specific to these underrepresented classes. Lastly, I also would like to point out that we must be sure that the annotated subclasses are correctly annotated during the data collection steps. For example, when we look at the confusion matrices, subclasses stage 2 and stage 3 were always predicted as subclass 4. This is very interesting. Here we should address the domain knowledge and make sure that the data correctly represent the subclasses.

**Comparison of model performances:**

**1-Complex model (with all features) versus simple model (with only selected features):** There are almost no differences between models. Both performed well with majority classes but didn’t perform well in minority classes. In this case, it is wise to choose the less complex model for the following reasons. Firstly, the principle of Occam's razor, which advocates for selecting the simplest explanation when multiple explanations are possible, aligns with the choice of the less complex model. The simple model is easier to explain. For example, it is easier to explain the underlying biological mechanisms of the simpler model. Additionally, its simplicity is very helpful where there is limited computational capacity or a preference for models that are easier to understand and maintain.

**2- Different classifiers:** Based on the evaluation of ROC-AUC and PR-AUC scores, it is evident that the Radial Kernel SVM exhibited the highest performance, closely followed by Logistic Regression, while Random Forest demonstrated relatively lower performance.

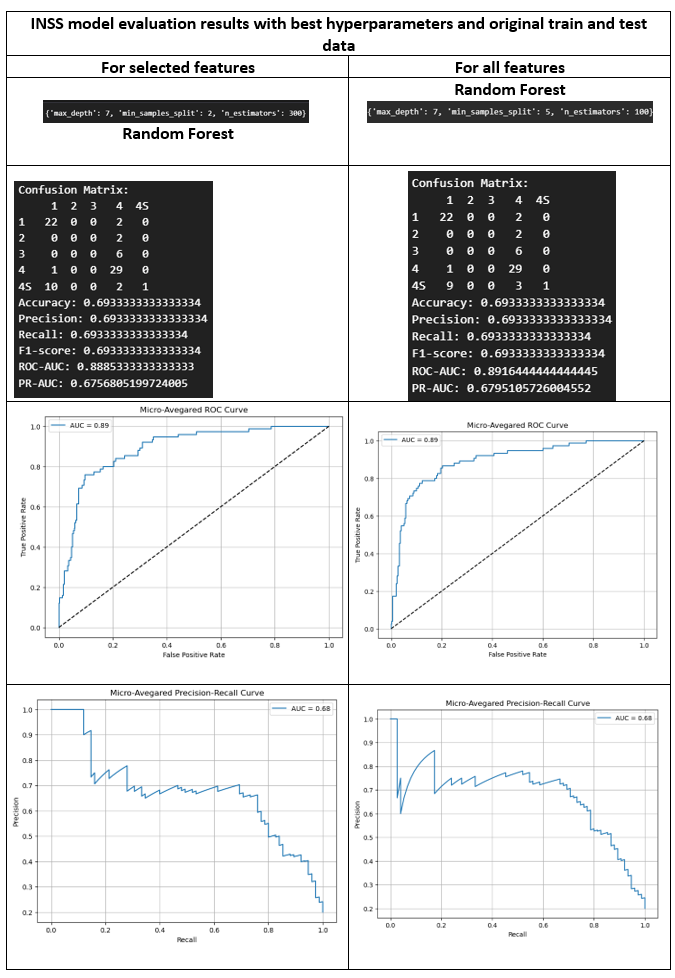
****

Figure 71: INSS.Stage model evaluation for Random Forest -for selected and all features matrices

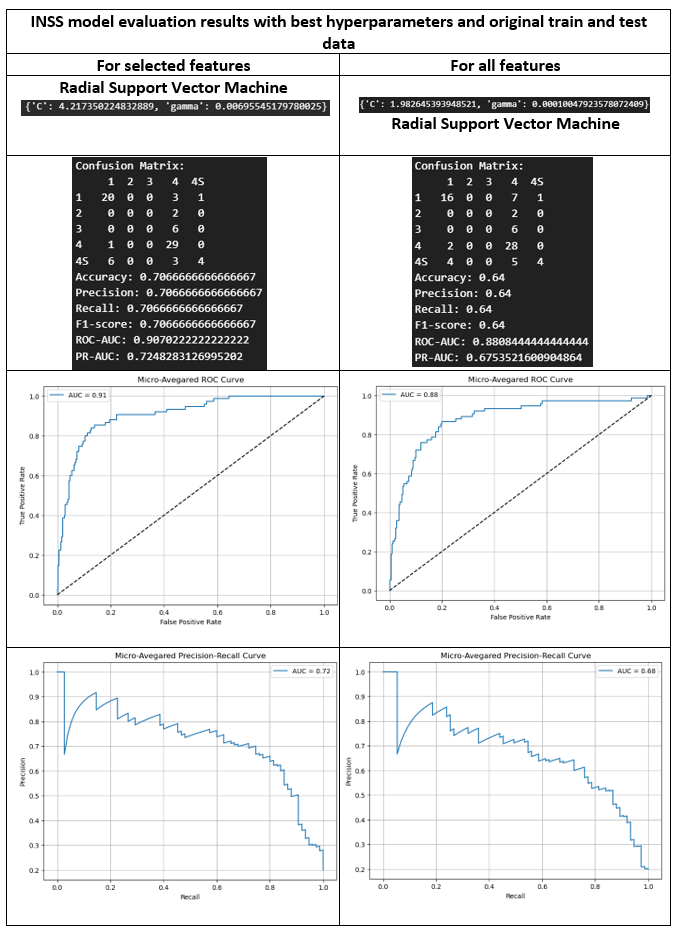


Figure 72: INSS.Stage model evaluation for SVM -for selected and all features matrices

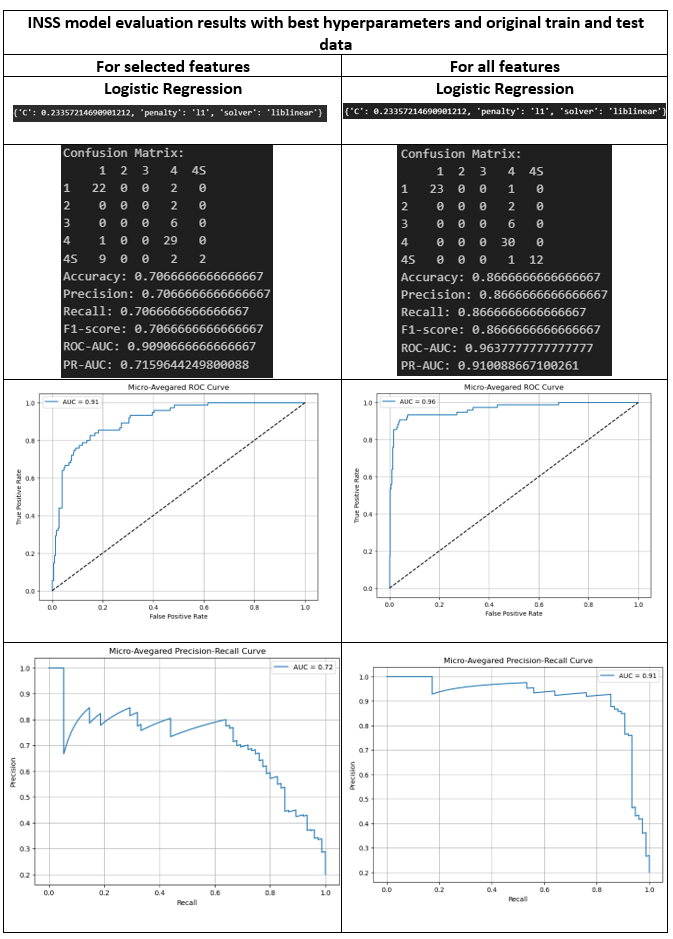


Figure 73 : INSS.Stage model evaluation for Logistic Regression -for selected and all features matrices

**6.2 HighRisk**

The HighRisk clinical endpoint stands out as the only target variable with balanced subclass distributions (see the EDA section). When evaluating models for HighRisk, these balanced sample classes have a significant positive impact on model performance, resulting in consistently high accuracy and precision scores (Figure 74-76). Furthermore, the use of proper feature selection techniques facilitated the model algorithms (classifiers) in accurately identifying patterns among the subclasses. The presence of balanced sample classes and effective feature selection along with the good hyperparameter choices collectively contribute to the exceptional performance of these models.

**Comparison of model performances:**

**1-Complex model (with all features) versus simple model (with only selected features):** In general**, less complex model performed well expect in Logistic Regression, complex model performed slightly better than less complex model. In this case, for the reasons explained in INSS.Stage part, less complex model selection will be more suitable choice.**

**2- Different classifiers:** After a comprehensive evaluation of ROC-AUC and PR-AUC scores, as well as a careful examination of the confusion matrices, it is evident that among all the classifiers, Logistic Regression emerged as the top performer. Following closely in second place was Random Forest, while SVM, although respectable, took the third position. This ranking is based on the models' overall performance in distinguishing between classes and highlights the strengths of each classifier in the context of the task.

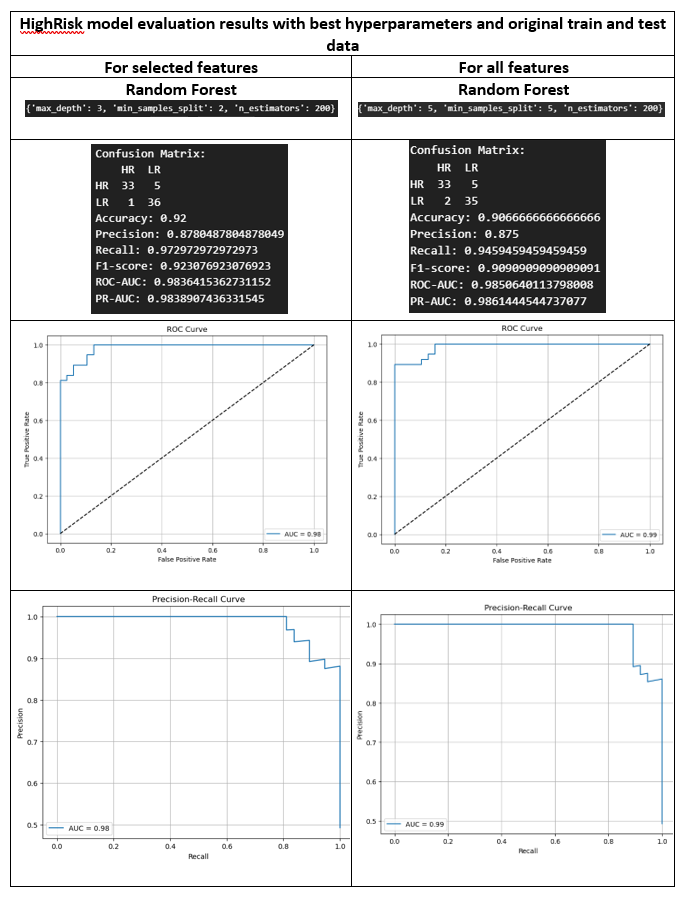
****

Figure 74 : HighRisk model evaluation for Random Forest -for selected and all features matrices

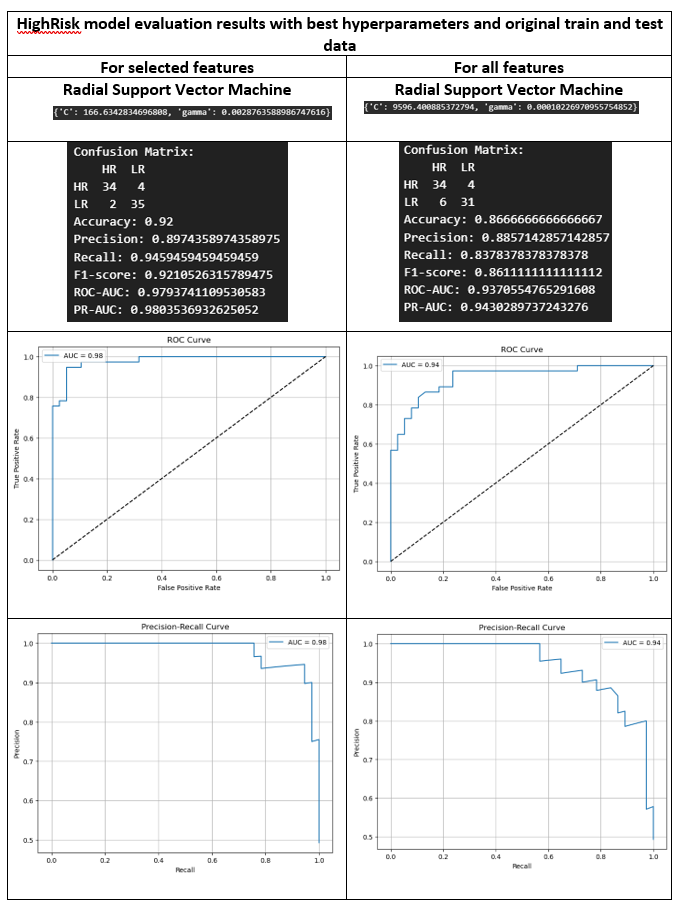
****

Figure 75 : HighRisk model evaluation for SVM -for selected and all features matrices

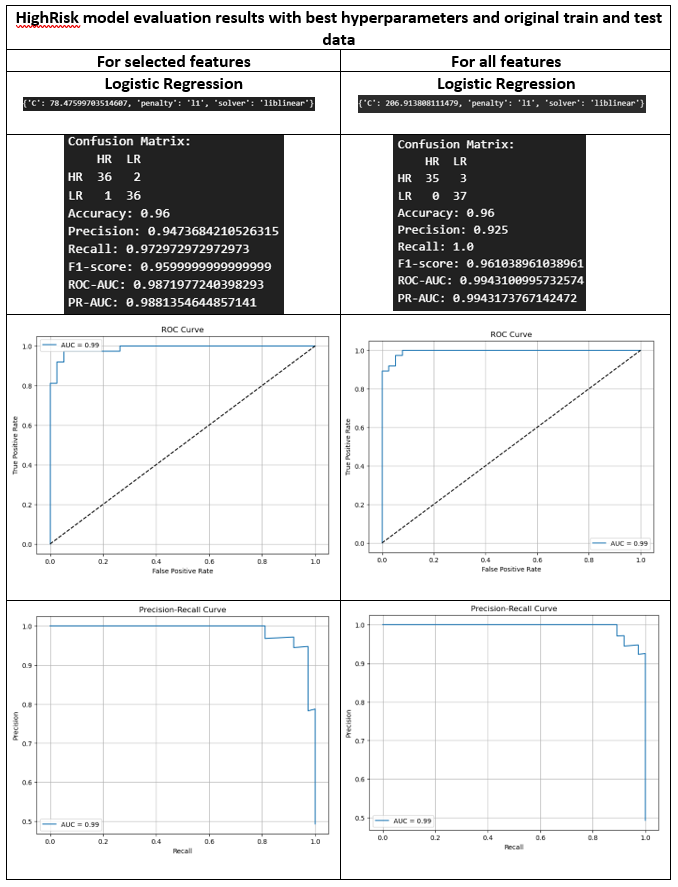
****

Figure 76 : HighRisk model evaluation for Logistic Regression -for selected and all features matrices

**6.3 Progression**

Much like the INSS.Stages clinical endpoint, the Progression clinical endpoint also exhibited imbalanced subclass distributions. However, the performance challenges observed were not limited to minority classes but extended to the majority classes as well (Figure 77-79). In contrast to the HighRisk clinical endpoint, the classifiers struggled to achieve the same level of performance.

This scenario raises two key considerations. Firstly, the selected features, or perhaps the overall feature matrix, may not be conducive to effectively discerning patterns within the two subgroups of this clinical endpoint. Secondly, a more in-depth exploration of hyperparameter tuning may be required. The effectiveness of hyperparameters plays a pivotal role in determining how models learn from the training data, and further tuning could potentially lead to improved results.

**Comparison of model performances:**

**1-Complex model (with all features) versus simple model (with only selected features):** In general**, less complex model performed much better. Especially Radial kernel SVM and Logistic regression performed very poorly with complex feature matrix.**

**2- Different classifiers:** In terms of ROC-AUC scores, Random Forest emerged as the top performer, with Logistic Regression following closely in second place, and Radial Kernel SVM ranking third. However, when considering PR-AUC scores, it was the Radial Kernel SVM that demonstrated the most impressive performance, securing the top spot. Random Forest ranked second in PR-AUC, while Logistic Regression held the third position. These distinctions in performance metrics underscore the varying strengths of each classifier under different evaluation criteria.

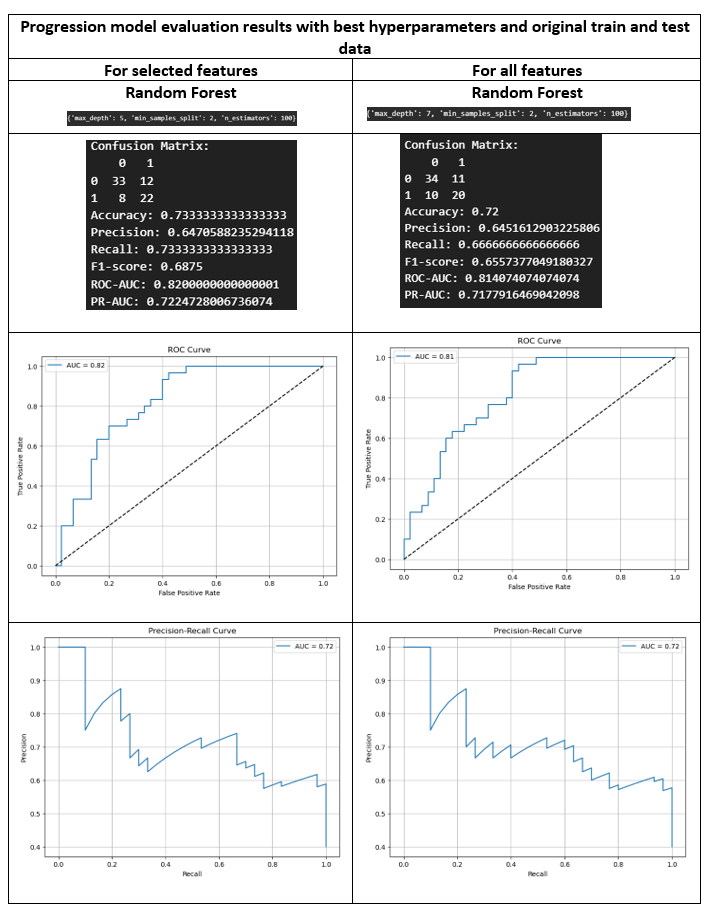
****

Figure 77 : Progression model evaluation for Random Forest -for selected and all features matrices

****

Figure 78 : Progression model evaluation for SVM -for selected and all features matrices

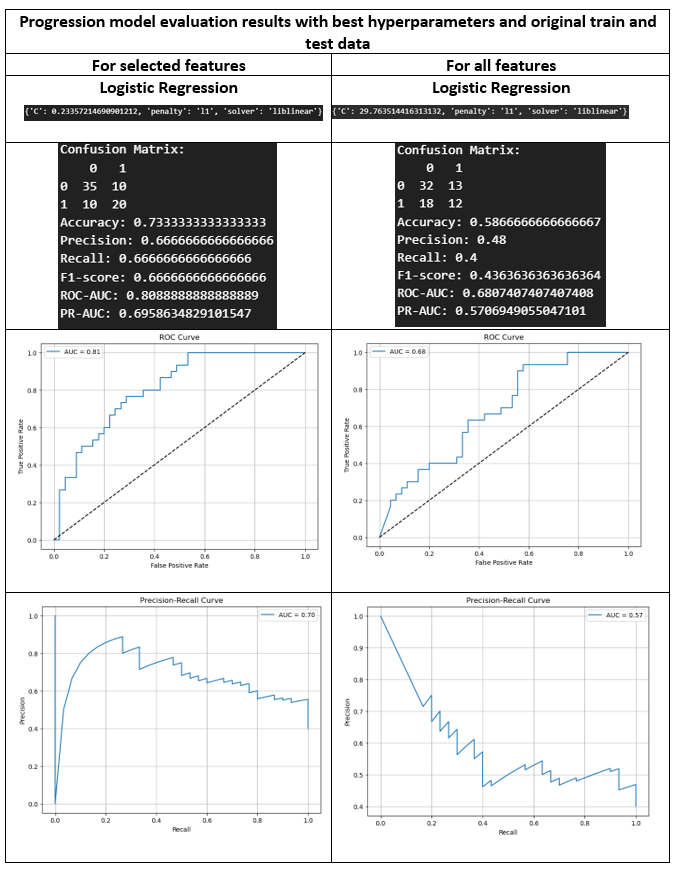
****

Figure 79 : Progression model evaluation for Logistic regression -for selected and all features matrices

**6.4 DeathFromDisease**

Much like INSS.Stages and Progression, the 'DeathFromDisease' clinical endpoint exhibited highly imbalanced subclass distributions. Among all three clinical endpoints, classifiers faced their most challenging task when dealing with this particular endpoint (Figure 80-82). However, reminiscent of the challenges seen in the Progression endpoint, these performance struggles extended beyond the minority classes and also affected the majority classes.

In stark contrast to the HighRisk clinical endpoint, where classifiers excelled, the struggle to achieve consistent performance across all three clinical endpoints was evident. As mentioned earlier, alongside the challenges posed by imbalanced class distributions and the necessity for more refined hyperparameter tuning, feature selection and hyperparameter configuration may also play significant roles in this performance gap. It's possible that the feature matrix, as it stands, wasn't able to effectively distinguish between the two classes.

**Comparison of model performances:**

**1-Complex model (with all features) versus simple model (with only selected features):** Once again, in general, less complex models outperformed their more intricate counterparts, with an interesting exception—Logistic Regression exhibited slightly superior results in this context.

**2- Different classifiers:** Evaluating the performance of different classifiers based on ROC-AUC and PR-AUC results, Random Forest emerged as the slightly better performer compared to Logistic Regression and Radial Kernel SVM.

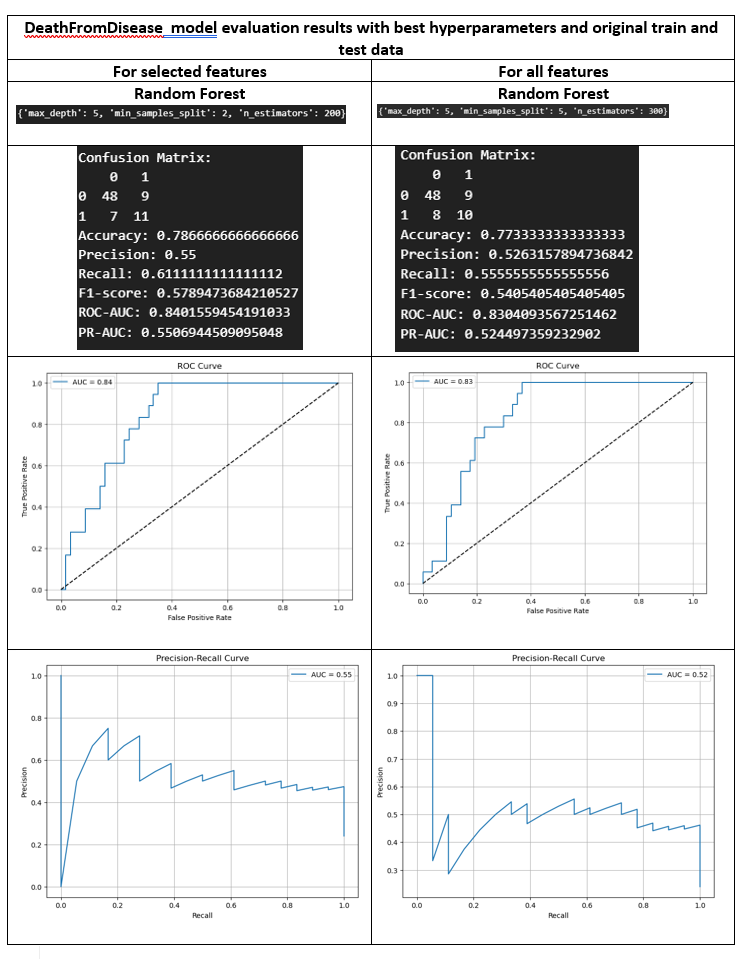
****

Figure 80 : DeathFromDisease model evaluation for RandomForest -for selected and all features matrices

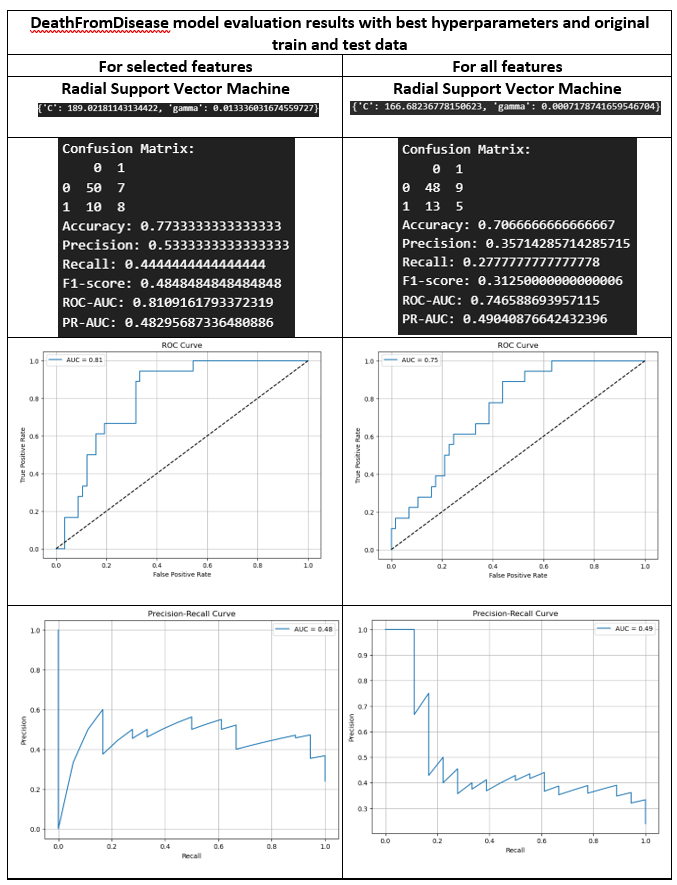
****

Figure 81 : DeathFromDisease model evaluation for SVM -for selected and all features matrices

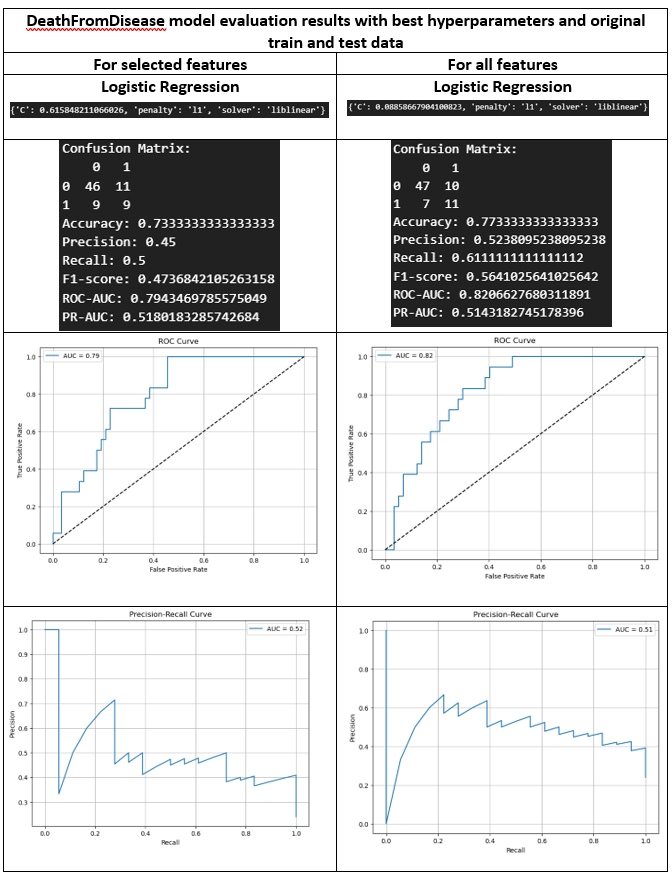
****

Figure 82 : DeathFromDisease model evaluation for Logistic Regression -for selected and all features matrices

**DISCUSSION**

This comprehensive study focused on predicting clinical endpoints in the context of neuroblastoma, a challenging pediatric cancer. By harnessing microarray data and essential clinical features, such as MYCN.status and clinic.genetic.subgroups, predictive models were developed and evaluated using Random Forest, Radial Kernel SVM, and Logistic Regression. Both complete and selected feature matrices were considered, resulting in a total of 24 predictive models.

This study revealed that predicting the HighRisk clinical endpoint exhibited outstanding model performance, thanks to balanced subclass distributions and effective feature selection. In contrast, the other clinical endpoints, such as INSS.Stage, Progression, and DeathFromDisease, faced challenges due to imbalanced subclasses, leading to less promising results, especially for minority classes. These findings underscore the critical importance of class balance in achieving robust clinical endpoint predictions. The success of HighRisk models highlights the advantages of having balanced data for accurate modeling, underlining the significance of collecting well-distributed samples. Moreover, it's noteworthy that among the various predictive models, less complex models built with selected features consistently demonstrated competitive performance with complex model built with all features, underscoring the importance of simplicity in modeling. Additionally, this study also demonstrates that careful feature selection and hyperparameter tuning are essential for achieving optimal model performance in clinical endpoint prediction.

Simplicity in machine learning models provides critical advantages. Firstly, simpler models are often more interpretable for clinicians or researchers to understand and trust these predictions. Complex models, on the other hand, may be seen as "black boxes" that provide accurate predictions but offer little insight into the underlying factors driving those predictions. Second, less complex models are computationally efficient. In a clinical setting, simpler models can provide rapid predictions, allowing for timely decision-making and deploy more be feasible or practical in a clinical environment in clinical environments where computational sources are limited.

In clinical research, it is often a problem to face imbalanced dataset which number of patient with (rare) condition, such as number of cancer patients, is often naturally underrepresented. For clinical endpoints characterized by imbalanced subclasses, there are possibilities for enhanced predictive accuracy. First, imbalanced data handling techniques are vital to address class imbalance. These strategies, such as oversampling the minority class (SMOTE) or undersampling the majority class (Tomek Link) or their balanced combinations (SMOTE+Tomek Links), contribute to a more balanced and representative dataset, which is crucial for reliable model outcomes (Khalilia et al., 2011, Mohammed et al., 2020, Kumar and Aggarwal, 2011, Taft et al., 2009, Wibowo and Fatichah, 2022). Second, ensuring accurate annotation of subclasses during data collection is paramount. This safeguards against potential misclassifications, ultimately enhancing the quality and integrity of the dataset. These collective efforts aim to overcome the challenges posed by imbalanced subclasses, ultimately leading to more robust and accurate clinical endpoint predictions. Third, feature engineering plays a central role, with comprehensive investigations into techniques that can tailor features to the specific characteristics of imbalanced subclasses, thereby yielding more accurate predictions. Here domain knowledge becomes critical. By leveraging subject-matter expertise to identify pertinent features and relationships within the data, domain knowledge can lead to more accurate predictions.

**CONCLUSION**

In summary, this study provides valuable insights into the challenges and opportunities in predicting clinical endpoints in the context of neuroblastoma, offering a foundation for future research to improve model performance and advance our understanding of this complex pediatric cancer. The exceptional performance of the HighRisk model highlights the need for more balanced datasets and the practical advantages of simplicity in modeling, underscoring the significance of these factors in clinical research and healthcare decision support. The ongoing exploration of feature engineering, imbalanced data handling techniques, and domain knowledge refinement promises to further enhance the predictive accuracy for challenging clinical endpoints.

**REFERENCES**

ACHARJEE, A., LARKMAN, J., XU, Y., CARDOSO, V. R. & GKOUTOS, G. V. 2020. A random forest based biomarker discovery and power analysis framework for diagnostics research. *BMC Medical Genomics,* 13**,** 178.

AKIBA, T., SANO, S., YANASE, T., OHTA, T. & KOYAMA, M. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* Anchorage, AK, USA: Association for Computing Machinery.

CASTEL, V., GARCı́A-MIGUEL, P., CAÑETE, A., MELERO, C., NAVAJAS, A., RUı́Z-JIMÉNEZ, J. I., NAVARRO, S. & BADAL, M. D. 1999. Prospective evaluation of the International Neuroblastoma Staging System (INSS) and the International Neuroblastoma Response Criteria (INRC) in a multicentre setting. *European Journal of Cancer,* 35**,** 606-611.

DAI, B., CHEN, R.-C., ZHU, S.-Z. & ZHANG, W.-W. Using random forest algorithm for breast cancer diagnosis. 2018 International symposium on computer, consumer and control (IS3C), 2018. IEEE, 449-452.

GAINS, J., MANDEVILLE, H., CORK, N., BROCK, P. & GAZE, M. 2012. Ten challenges in the management of neuroblastoma. *Future Oncology,* 8**,** 839-858.

HUANG, S., CAI, N., PACHECO, P. P., NARRANDES, S., WANG, Y. & XU, W. 2018. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genomics Proteomics,* 15**,** 41-51.

IQBAL, M. J., JAVED, Z., SADIA, H., QURESHI, I. A., IRSHAD, A., AHMED, R., MALIK, K., RAZA, S., ABBAS, A., PEZZANI, R. & SHARIFI-RAD, J. 2021. Clinical applications of artificial intelligence and machine learning in cancer diagnosis: looking into the future. *Cancer Cell International,* 21**,** 270.

KAGGLE Evaluation Metrics For Multi-class Classification.

KHAIRUNNAHAR, L., HASIB, M. A., REZANUR, R. H. B., ISLAM, M. R. & HOSAIN, M. K. 2019. Classification of malignant and benign tissue with logistic regression. *Informatics in Medicine Unlocked,* 16**,** 100189.

KHALILIA, M., CHAKRABORTY, S. & POPESCU, M. 2011. Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making,* 11**,** 51.

KONONENKO, I. 2001. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine,* 23**,** 89-109.

KUMAR, K. & AGGARWAL, R. 2011. IMBALANCED DATASET CLASSIFICATION AND SOLUTIONS: A REVIEW. *Int. J. Comput. Bus. Res,* 2.

MATTHAY, K. K., MARIS, J. M., SCHLEIERMACHER, G., NAKAGAWARA, A., MACKALL, C. L., DILLER, L. & WEISS, W. A. 2016. Neuroblastoma. *Nature Reviews Disease Primers,* 2**,** 16078.

MERSEBURGER, A. S., ANASTASIADIS, A. G., HENNENLOTTER, J., SCHILLING, D., SIMON, P., MACHTENS, S. A., SERTH, J., STENZL, A. & KUCZYK, M. A. 2006. Tissue microarrays: applications in urological cancer research. *World Journal of Urology,* 24**,** 579-584.

MOHAMMED, A. J., HASSAN, M. M. & KADIR, D. H. 2020. Improving classification performance for a novel imbalanced medical dataset using SMOTE method. *International Journal of Advanced Trends in Computer Science and Engineering,* 9**,** 3161-3172.

MORGENSTERN, D. A., BARUCHEL, S. & IRWIN, M. S. 2013. Current and Future Strategies for Relapsed Neuroblastoma: Challenges on the Road to Precision Therapy. *Journal of Pediatric Hematology/Oncology,* 35**,** 337-347.

OHIRA, M., OBA, S., NAKAMURA, Y., HIRATA, T., ISHII, S. & NAKAGAWARA, A. 2005a. A review of DNA microarray analysis of human neuroblastomas. *Cancer Letters,* 228**,** 5-11.

OHIRA, M., OBA, S., NAKAMURA, Y., ISOGAI, E., KANEKO, S., NAKAGAWA, A., HIRATA, T., KUBO, H., GOTO, T., YAMADA, S., YOSHIDA, Y., FUCHIOKA, M., ISHII, S. & NAKAGAWARA, A. 2005b. Expression profiling using a tumor-specific cDNA microarray predicts the prognosis of intermediate risk neuroblastomas. *Cancer Cell,* 7**,** 337-50.

OTTE, J., DYBERG, C., PEPICH, A. & JOHNSEN, J. I. 2021. MYCN function in neuroblastoma development. *Frontiers in oncology,* 10**,** 624079.

QIAN, X., ZHOU, Z., HU, J., ZHU, J., HUANG, H. & DAI, Y. 2021. A comparative study of kernel-based vector machines with probabilistic outputs for medical diagnosis. *Biocybernetics and Biomedical Engineering,* 41**,** 1486-1504.

SABERI-KARIMIAN, M., KHORASANCHI, Z., GHAZIZADEH, H., TAYEFI, M., SAFFAR, S., FERNS, G. A. & GHAYOUR-MOBARHAN, M. 2021. Potential value and impact of data mining and machine learning in clinical diagnostics. *Critical Reviews in Clinical Laboratory Sciences,* 58**,** 275-296.

SHAMAI, G., BINENBAUM, Y., SLOSSBERG, R., DUEK, I., GIL, Z. & KIMMEL, R. 2019. Artificial Intelligence Algorithms to Assess Hormonal Status From Tissue Microarrays in Patients With Breast Cancer. *JAMA Network Open,* 2**,** e197700-e197700.

SOCIETY, A. C. *Neuroblastoma Stages and Prognostic Markers* [Online]. Available: <https://www.cancer.org/cancer/types/neuroblastoma/detection-diagnosis-staging/staging.html> [Accessed 2023].

STACKEXCHANGE Do we normalise the dataset before or after performing one hot encoding?

SWANSON, K., WU, E., ZHANG, A., ALIZADEH, A. A. & ZOU, J. 2023. From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. *Cell,* 186**,** 1772-1791.

TAFT, L. M., EVANS, R. S., SHYU, C. R., EGGER, M. J., CHAWLA, N., MITCHELL, J. A., THORNTON, S. N., BRAY, B. & VARNER, M. 2009. Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery. *Journal of Biomedical Informatics,* 42**,** 356-364.

WIBOWO, P. & FATICHAH, C. 2022. Pruning-based oversampling technique with smoothed bootstrap resampling for imbalanced clinical dataset of Covid-19. *Journal of King Saud University - Computer and Information Sciences,* 34**,** 7830-7839.

ZANGMO, C. & TIENSUWAN, M. Application of logistic regression models to cancer patients: a case study of data from Jigme Dorji Wangchuck National Referral Hospital (JDWNRH) in Bhutan. Journal of Physics: Conference Series, 2018. IOP Publishing, 012031.

ZHANG, W., YU, Y., HERTWIG, F., THIERRY-MIEG, J., ZHANG, W., THIERRY-MIEG, D., WANG, J., FURLANELLO, C., DEVANARAYAN, V., CHENG, J., DENG, Y., HERO, B., HONG, H., JIA, M., LI, L., LIN, S. M., NIKOLSKY, Y., OBERTHUER, A., QING, T., SU, Z., VOLLAND, R., WANG, C., WANG, M. D., AI, J., ALBANESE, D., ASGHARZADEH, S., AVIGAD, S., BAO, W., BESSARABOVA, M., BRILLIANT, M. H., BRORS, B., CHIERICI, M., CHU, T.-M., ZHANG, J., GRUNDY, R. G., HE, M. M., HEBBRING, S., KAUFMAN, H. L., LABABIDI, S., LANCASHIRE, L. J., LI, Y., LU, X. X., LUO, H., MA, X., NING, B., NOGUERA, R., PEIFER, M., PHAN, J. H., ROELS, F., ROSSWOG, C., SHAO, S., SHEN, J., THEISSEN, J., TONINI, G. P., VANDESOMPELE, J., WU, P.-Y., XIAO, W., XU, J., XU, W., XUAN, J., YANG, Y., YE, Z., DONG, Z., ZHANG, K. K., YIN, Y., ZHAO, C., ZHENG, Y., WOLFINGER, R. D., SHI, T., MALKAS, L. H., BERTHOLD, F., WANG, J., TONG, W., SHI, L., PENG, Z. & FISCHER, M. 2015. Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biology,* 16**,** 133.