Applying deep learning to integrate multi-omics data for survival analysis in non-small cell lung cancer patients

Bach Tran
University of Colorado Boulder
bach.tran-1@coloardo.edu

ABSTRACT

Non-small cell lung cancer (NSCLC) is the most prevalent form of lung cancer, and it remains a leading cause of cancer mortality worldwide. NSCLC patient outcomes highly depend on multiple factors including demographics, lifestyles, and genetics. Although advancements in high-throughput sequencing have made patients' molecular profiles accessible, the prognosis of NSCLC patients based on molecular data is often challenging due to complex cancer biology. The recent development of deep learning approaches for survival analysis can offer new opportunities to improve survival predictions from patients' sequencing data. In this project, I propose a deep-learning framework that integrates clinical, DNA methylation, gene expression, and somatic mutation data to predict patient survival. Univariate Cox proportional hazards regression was applied for feature selection within each omics type to identify important molecular biomarkers for NSCLC patient survival. Selected features were used to optimize separate deep survival networks for each omics dataset. The trained models were integrated through ensemble techniques, including averaging, weighted averaging, and stacking. By combining the predictive strengths of each model, I developed a comprehensive survival analysis model that can accurately predict patient outcomes and identify key molecular factors associated with NSCLC prognosis. The ensemble models significantly outperformed the individual models, with the stacking ensemble being the best-performing model. This approach improved accuracy and generalizability in survival prediction, helping personalized treatment strategies for NSCLC patients.

INTRODUCTION

Lung cancer is responsible for a significant proportion of global cancer cases and deaths. According to data from the Global Cancer Observatory, in 2020, it was estimated there were 2.2 million new lung cancer diagnoses, causing 1.8 million deaths [1]. This high incidence and mortality rate make lung cancer the leading cause of cancer death worldwide, highlighting the need for accurate and effective prognostic models to improve clinical outcomes. Non-small cell lung cancer (NSCLC) accounts for approximately 84% of all lung cancers [2]. Despite advances in medical interventions, NSCLC patients have a low five-year survival rate due to late-stage diagnosis. NSCLC patient outcomes can vary depending on molecular and clinical factors. This underscores the need for a comprehensive prognostic tool that can predict NSCLC patient survival with high accuracy.

Multiple omics data, including transcriptomics, DNA methylation, and mutation profiles, can provide insights into disease progression and prognosis. Each sequencing data type captures a unique aspect of cancer biology. DNA methylation, an epigenetic modification that involves adding a methyl group to DNA molecules, primarily at CpG sites, plays a key role in regulating gene expression. Methylation levels are known to be closely related to the expression of tumor-related genes, showing the role of DNA methylation in cancer biology [3]. Gene expression profiling, typically through RNA sequencing, provides a snapshot of the transcriptional activity in cancerous cells, reflecting the functions of genes within the tumor. These profiles can shed light on the pathways and molecular signatures associated with patient outcomes and tumor aggressiveness. Mutation data captures genetic alterations that contribute to oncogenesis and influence cancer progression. Specific mutations in genes such as EGFR, KRAS, and TP53 are known to affect NSCLC prognosis and response to treatment [4]. Each data type contains valuable information about disease pathology and has been associated with NSCLC patient outcomes. Most previous studies have focused on using individual data types to predict patient survival. However, using only one type of omics data often leads to an incomplete prediction of the patient outcome due to the complex, multi-layered nature of cancer biology. Although these studies have promising results and shed light on the disease biology of NSCLC, integrating these data types can offer a more comprehensive approach to survival analysis as it allows researchers to identify various biological interactions influencing patient outcomes.

When applied to multi-omics data, traditional survival analysis models often struggle with high dimensionality and collinearity. The most widely used model for survival analysis, the Cox proportional hazards (CPH) model, is limited by the assumption of linearity and proportional hazards, which do not always hold in biological data. Moreover, high dimensionality and collinearity are common in multi-omics data, in which a single dataset can include tens of thousands of features. This causes the models to overfit and have low performance. The CPH model also suffers from collinearity due to the non-linear relationships and interactions between features. For instance, methylation levels are often negatively correlated to gene expression. Alternative

machine learning approaches, such as random survival forests (RSF), have been introduced to handle data with higher dimensions, but they often lack the flexibility to capture complex biological interactions between different omics datasets.

Deep survival networks provide a more flexible and effective approach for handling high-dimensional, non-linear data like multi-omics datasets. DeepSurv, introduced recently by Katzman et al., utilizes the Cox partial likelihood loss function to model the log-risk hazard of an event [5]. DeepSurv can model non-linear relationships, making it particularly well-suited to multi-omics integration, as it can account for complex interactions across genetic, epigenetic, and transcriptomic data layers. Moreover, DeepSurv also offers a customizable deep-learning architecture with fully connected dense layers to predict the hazard rate, allowing researchers to have flexibility when modeling high-dimensional data. However, to my knowledge, no previous studies have successfully integrated deep learning and multi-omics data specifically for NSCLC prognosis.

In this project, I aim to develop a deep learning-based framework based on DeepSurv that integrates DNA methylation, gene expression, and mutation data for improved survival analysis in NSCLC patients. I propose to build separate deep survival networks for each omics type and combine them via ensemble and stacking methods. The integration of multi-omics data using deep learning methods presents a promising direction in NSCLC survival analysis. By combining DNA methylation, gene expression, and mutation data in a single framework, my project hopes to produce a more comprehensive and accurate prognostic model that better reflects the complexity of NSCLC biology. This approach could lead to new insights into the molecular drivers of NSCLC, inform treatment decisions, and ultimately contribute to improving survival rates in this high-risk patient population.

RELATED WORK

Since its introduction in 2018, DeepSurv has been used in multiple studies to predict disease risk as well as patient survival in multiple diseases. Kim et al. applied the model to predict the overall survival of oral cancer patients based on clinical factors and achieved a high c-index of 0.8 in the testing set [6]. The study compared the performance of DeepSurv to the CPH model and RFS, showing the superior performance of DeepSurv with the dataset. Another recent study used the deep survival model to predict the risk of diabetes mellitus, hypertension, and dyslipidemia over a 7-year period [7]. This study suggests that DeepSurv can generalize the data and capture patterns well, resulting in higher performance compared to CPH. Studies like these emphasize the potential of using deep neural networks tailored to specific data structures to enhance survival prediction. DeepSurv has since been applied to various cancer datasets, outperforming traditional survival models in accuracy and generalizability.

Multi-omics integration has emerged as a promising approach for cancer prognosis as it can provide a comprehensive understanding of disease pathology. Recent studies have shown that multi-omics data integration can improve prognostic models for diseases like breast and lung cancer by identifying novel interactions between molecular layers. Malik et al. integrated gene expression, mutation, methylation, protein level, and other clinical features using deep learning models and predicted the survival of breast cancer patients with an accuracy of 94% [8]. Their model was also able to predict drug response with a high performance. The authors successfully employed late integration methods, in which separate deep learning models are trained on each omics dataset and subsequently combined with stacking or ensemble. This is the approach I proposed to apply to NSCLC patients. However, the research treated survival as a binary classification without taking the survival period and hazard estimation into consideration. My project built on these insights and integrates multiple deep survival networks through ensemble techniques to capture the complex interplay between methylation, gene expression, and mutation data. This approach would provide more accurate and biologically informative predictions for NSCLC prognosis, contributing to the ongoing development of personalized oncology.

PROPOSED WORK

I proposed an analysis pipeline that includes data collection and preprocessing for each omics data type, feature selection using CPH, and building deep survival models. The following proposed work will show a detailed plan for my analyses.

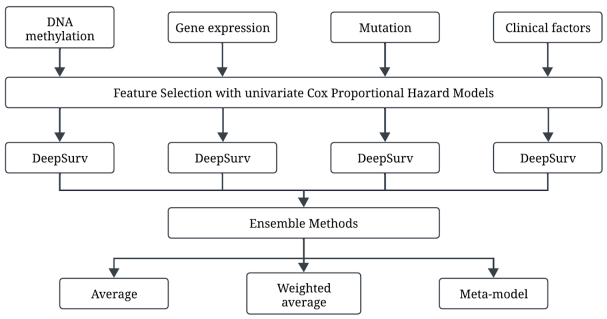

Figure 1: Proposed analysis pipeline. The preprocessed data will be modeled with univariate Cox Proportional Hazard Models for feature selection. The selected features will be used to optimize individual DeepSurv models, which will be integrated via ensemble methods including average, weighted average, and meta-model.

1 Data Collection and Preprocessing

Patient data were obtained from The Cancer Genome Atlas (TCGA) dataset for Lung Adenocarcinoma (LUAD). TGCA datasets are deidentified and publicly available for any use. The LUAD dataset includes clinical factors, such as age, sex, race, and tumor stage, as well as DNA methylation, RNA sequencing, and somatic mutation data. This dataset contains data on 566 patients with NSCLC.

Different data types required a special preprocessing method. Clinical data contain information on patients' ages, sexes, genetic ancestries, overall tumor stages, primary tumor sizes, number of nearby lymph nodes, metastasis statuses, and overall survival data. Patients with missing survival status and duration were removed from the study. Categorical features were encoded as binary or ordinal features. The number of missing data in all features is relatively small (<10%). Thus, missing numerical data were imputed with the median, while missing categorical data were imputed with the mode.

RNA sequencing data were imported as read counts for each gene. Genes with low mean read counts (<15) were removed to avoid redundant data. The read counts were log-transformed and standardized to z score so that gene expression levels have a normal distribution with a mean of 0. I also filtered out features with variance lower than 0.05 due to their low predictive values. After preprocessing, the RNA sequencing data has 19,414 remaining genes.

In the TCGA dataset, methylation data was already normalized to methylation beta values and scaled to range from 0 to 1. Missing values are more common in this dataset. Therefore, features with more than 20% missing values were removed. The remaining missing values were imputed using the K nearest neighbor (KNN) imputation method. This method was chosen because of its robustness in imputation datasets with a large number of features, and it is a common method for omics datasets. The processed DNA methylation data has 13,065 remaining features.

Mutation data record individual mutations, including missense, inframe, truncating, and splice, for each sample. To simplify the data, I converted mutation data to binary features representing Wild type and Mutation. This process helped generalize the mutation data because specific mutation types are not common among patients. Features with variants less than 0.01 were removed from the dataset because they do not offer any predictive values. After preprocessing, the mutation data has 10,483 remaining features. Only patients with available clinical, gene expression, methylation, and mutation data were selected for the study. This resulted in 499 patients for this study. After preprocessing, the different omics data were suitable for modeling.

2 Feature Selection Using Cox Proportional Hazards Model

Due to the high dimension of sequencing data which can have tens of thousands of features, feature selection is necessary before modeling. Despite having large dimensions, these data often have a large portion of features with low variances and low predictive values. To select important features of patient survival, I fitted each feature into a univariate CPH model. This model can handle both time and censorship features of survival studies, making it suitable for feature selection. The model can evaluate the effects of covariates on the rate of death at a particular time point using the hazard function described below.

$$\lambda(x) = \lambda_0(t) \bullet e^{h(x)} \quad (1)$$

$$h(x) = b_1 x_1 + b_2 x_2 + \ldots + b_p x_p \quad (2)$$

The model assumes the hazard function is the product of a baseline hazard function, $\lambda_0(t)$, and a risk score, $e^{h(x)}$. The risk score defines the effect of covariates on patient survival via the log-risk function $h(x)$, which is a linear function of patient covariates. CPH can be optimized using the Cox partial likelihood. This likelihood is defined with the following formula with parameterized weights β.

$$L(\beta) = \prod i: e_i = \frac{e^{h\beta x_i}}{\sum j \in R(t_i) e^{h\beta x_i}} \quad (3)$$

The values t_i, e_i, and x_i denote event time, indicator, and baseline data for the ith observation respectively.

For each omics dataset, I found the coefficient and p-value of each feature. The p-values were corrected for multiple comparisons using the Benjamini-Hochberg False discovery rate method. Multiple testing corrections for omics data were necessary because the large number of features drastically increases the false negative rate. Features with a False discovery rate <0.05 were identified as significantly associated with survival outcomes and were selected for future analysis. I identified 65, 52, and 46 significant features for gene expression, methylation, and mutation data, respectively. After the feature selection process, the dimensions of each data type were reduced, focusing on the most prognostic markers. This process prevented deep learning models from the curse of dimensionality and improved the models' performance significantly.

3 DeepSurv Models

DeepSurv is a feed-forward neural network that predicts the effects of patients' features on their hazard functions. The model has the input of patients' covariates. The inputs are passed through multiple hidden fully connected layers with nodes and weights θ. Each fully connected layer is followed by a dropout layer. The output layer has a linear activation function that estimates the log-risk value $\hat{h(x)}$ in the Cox regression.

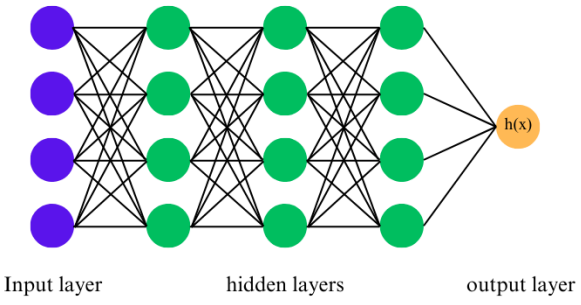

Figure 2: The overall model architecture of DeepSurv.

Weights θ are optimized using a loss function of negative log partial likelihood $L(\beta)$ from the Cox regression with an additional L2 regularization.

$$l(\theta) = -\frac{1}{N_{E=1}} \sum_{i:E_i=1} \left(\hat{h}_\theta(x_i) - log \sum_{j \in R(T_i)} e^{\hat{h}_\theta(x_i)} \right) + \lambda \|\theta\|_2^2 \quad (4)$$

This neural network estimation of the hazard function relaxes the linear assumption of the CPH Model. This allows DeepSurv to model nonlinear feature interactions in sequencing data without adding high-level interaction terms. Therefore, DeepSurv is more appropriate when modeling nonlinear omics data in my project.

For this project, I proposed to build and optimize four individual DeepSurv networks to model clinical, DNA methylation, gene expression, and mutation data. These models used the features important to patient survival from the previous feature selection process as inputs. The models utilized the negative log partial likelihood loss function to estimate the log-risk hazard values.

Individual DeepSurv models were optimized using the Adaptive Moment Estimation (Adam) gradient descent algorithm thanks to its robustness and fast convergence. Nesterov momentum, learning rate scheduling, and early stopping were also implemented during the training process to help the models converge and prevent the risk of overfitting. For the training process, the patient datasets were split into training and testing sets with a ratio of 80:20. The models were trained with a 5-fold cross-validation of the training set. Hyperparameters, including learning rate, decay rate, L2 regularization, batch size, and number of neurons and layers, were tuned using a customized function to perform 5-fold cross-validation training. The best hyperparameters were selected based on the model's accuracy when predicting the testing set. Individual DeepSurv models were trained over 500 epochs with an early stopping of 50 epochs after the model's validation loss did not decrease.

4 Ensemble Methods

After the optimization process, the 4 trained DeepSurv models were integrated through multiple ensemble approaches. The first method was averaging ensemble, in which the average of the output log-risk hazard estimations from 4 individual models is determined. This is a simple and effective method to combine the outputs of different models. However, the averaging ensemble method could reduce the overall accuracy if some models have lower performance than others.

Weighted average ensemble could address this problem by finding the weighted average of the model outputs based on the model performance. Models with a higher predictive accuracy have a larger weight on the average and vice versa.

Another ensemble approach was the stacking ensemble. This method uses a meta-learning algorithm, in which a meta-model is built by combining predictions from individual DeepSurv models to make a final output. The meta-model was a neural network with a small number of hidden layers and an output layer. The meta-model was trained and optimized similarly to the individual models. However, during the training process for the meta-model, the individual DeepSurv models' parameters were not changed. I tuned the ensemble methods to achieve the optimal combination of the individual models for improved survival prediction accuracy.

EVALUATION

1 Evaluation metrics

The accuracy of the deep Survival models was evaluated using the concordance index (c-index) and Kaplan-Meier (KM) methods. The most common evaluation metric in survival analysis is the c-index, which shows the model's ability to provide a correct ranking of survival times based on individual risk scores. The main idea behind the concordance score is that at a time, dying patients should have a higher risk score than the surviving patients. The c-index reflects how well a model predicts the ordering of patients' death times. A random model has a c-index of approximately 0.5, while a perfect model has a c-index of 1.

Another accuracy metric the models were evaluated on is KM, a non-parametric statistical tool used for estimating the survival function. This method generates a survival curve that visualizes the probability of an individual surviving past certain time points, even when data include censored observations. Each step in the Kaplan-Meier curve represents a survival event, with a decrease at each event time and a flat line in between indicating periods with no events. I used the models' prediction of patient risk scores to split the test data into high-risk and low-risk groups at the median. The KM curves for high-risk and low-risk groups were visualized. I also performed a log-rank test to compare the survival distribution of the groups. The log-rank test is often used with the KM method to assess whether there is a statistically significant difference between the survival curves. A high-performing model will result in a significant difference between the KM curves of the two groups.

To evaluate the DeepSurv models, the dataset was split into training and testing sets with a ratio of 80 to 20. The models were trained with a 5-fold cross-validation of the training set. The

predictive accuracy of models was evaluated on the unseen testing set. I performed bootstrapping and sampled the testing set with replacements for the c-index. This allowed me to perform two-tailed student t-tests to compare the performance of models pairwise. The ensemble models were compared to the individual models to determine whether integrating multi-omics data increased the model's accuracy. The different ensemble models were also compared pairwise to find the optimal ensemble method. The evaluation process determined if integrating datasets improved survival analysis predictive accuracy and which ensemble approach yielded the highest accuracy.

2 Results

2.1 Feature Selection

The selected patients for this study have an age range of 38 to 88, with a median of 67. This suggests that the study focuses on older patients. The patients' ages in my study also represent the overall NSCLC patient population, as NSCLC is known to be more common in the older age groups. Moreover, around 53% of the patients are female. This suggests that the sex feature in patients is balanced. However, the majority of the patients have genetic European ancestry, which implies that this study's findings should mostly be applied to NSCLC of European descent. The survival status of the patient population is not balanced, as the majority of the patients are alive. This imbalanced data could reduce predictive models' performance by increasing models' biases. To prevent the effect of imbalanced data, I split the training and testing sets using the stratified split method, ensuring the same proportion of living and deceased patients in datasets.

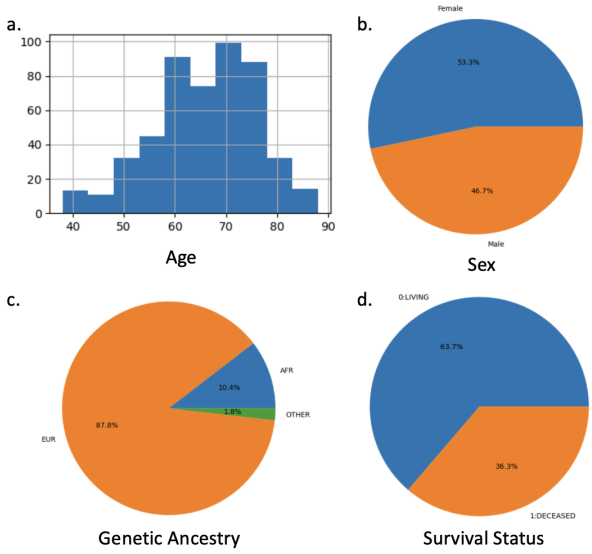

Figure 3: Patient demographics. a. Age; b. Sex; c. Genetic ancestry; d. Survival status.

I performed feature selection to the omics datasets to reduce the dimensions and filter only features with high predictive values. The features with a high number of missing values (>20%) or low variance were filtered. Each feature was modeled with a univariate CPH model to find its effect on overall survival. The p-values from CPH models were corrected with the Benjamini-Hochberg procedure. After feature selection, 65, 52, and 46 significant features were selected from gene expression, methylation, and mutation datasets.

	Gene expression	Methylation	Mutation
Original	20531	22601	243229
Filtered	19414	13065	10483
Selected	65	52	46

Table 1: The feature selection process.

2.2 Individual DeepSurv models

Individual DeepSurv models were used to predict patients' survival risk scores from clinical and sequencing data. The models were optimized and hyperparameters were tuned using a 5-fold cross-validation method. The DeepSurv for clinical data was trained over 500 epochs and an early stopping of 50 epochs. After optimization, the highest-performing model based on prediction accuracy in the test set had 2 fully connected hidden layers with 32 filters, a learning rate of 0.001, a decay rate of 0.001, and an L2 regularization penalty of 0.1. The optimized model achieved a mean training c-index of 0.71 (95% CI: 0.67-0.76) and a mean testing c-index of 0.67 (95% CI: 0.58-0.76). This can be considered a high accuracy, considering the model only has 8 features on patients' clinical information.

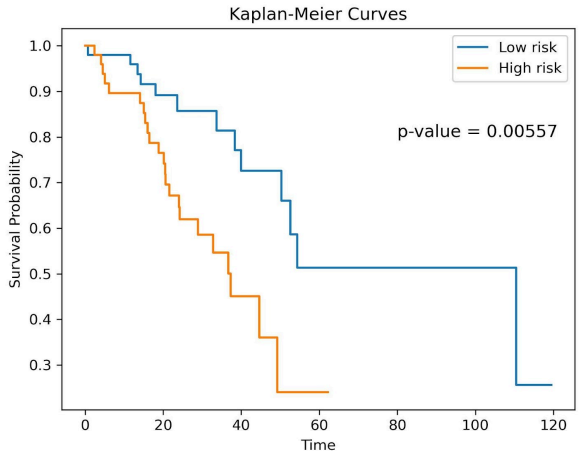

Figure 4: KM curves of the Clinical DeepSurv's risk score predictions in the testing set.

When applied to the unseen testing dataset, the model could predict patients' risk scores with high accuracy. I divided the testing patients into high-risk and low-risk groups based on the median of the predicted risk scores. According to the KM curves, the high-risk patients have a higher death rate and worse survival prognosis compared to the low-risk group. The log-rank test suggests this difference between the groups is statistically significant ($p<0.05$). This result indicates that the DeepSurv model can accurately predict patients' risk scores from clinical factors.

The RNAseq DeepSurv was also trained over 500 epochs and an early stopping of 50 epochs. After optimization, the highest-performing model had 3 fully connected hidden layers with 32 filters, a learning rate of 0.001, a decay rate of 0.001, and an L2 regularization penalty of 0.001. The trained model achieved a mean c-index of 0.92 (95% CI: 0.9-0.94) and 0.71 (95% CI: 0.58-0.80) respectively for the training and testing sets. There was an improvement in performance in RNAseq DeepSurv compared to clinical DeepSurv.

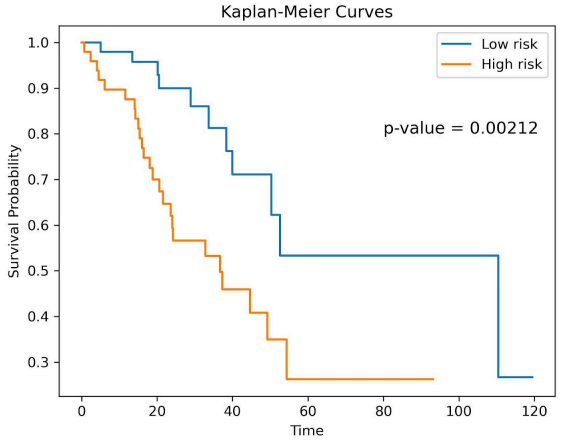

Figure 5: KM curves of the RNAseq DeepSurv's risk score predictions in the testing set.

The model accurately predicted patients' risk scores in the unseen testing set. The KM survival curves revealed that patients in the high-risk group had a higher mortality rate and poorer survival outcomes compared to those in the low-risk group. The log-rank test confirmed that this difference between the groups was statistically significant ($p<0.05$). These findings indicate that the DeepSurv model effectively predicts patient risk scores using RNA sequencing data.

The Methylation DeepSurv was optimized with 3 fully connected hidden layers with 64 filters, a learning rate of 0.001, a decay rate of 0.001, and an L2 regularization penalty of 0.01. The trained model achieved a mean c-index of 0.92 (95% CI: 0.9-0.94) and 0.68 (95% CI: 0.59-0.78) respectively for the training and testing sets.

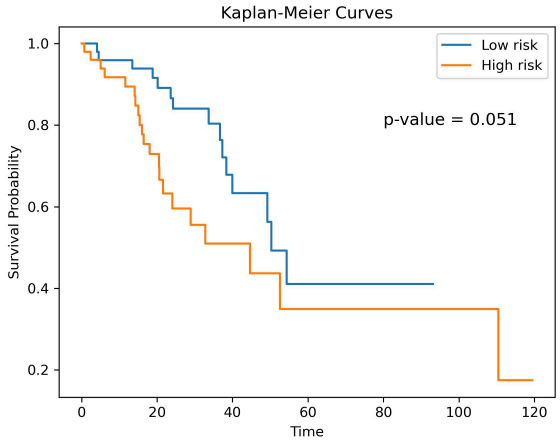

Figure 6: KM curves of the Methylation DeepSurv's risk score predictions in the testing set.

The methylation model also accurately separated high-risk and low-risk patients. The KM curves and the log-rank test showed a significant difference between the survival of the predicted patient groups ($p<0.05$). This suggests that the DeepSurv model could predict patients' risk scores in the testing set with high accuracy from methylation data.

The Mutation DeepSurv was also trained over 500 epochs and an early stopping of 50 epochs. The best-performing model had 2 fully connected hidden layers with 64 filters, a learning rate of 0.01, a decay rate of 0.001, and an L2 regularization penalty of 0.1. The optimized model achieved a mean c-index of 0.81 (95% CI: 0.78-0.85) and 0.74 (95% CI: 0.67-0.82) respectively for the training and testing sets.

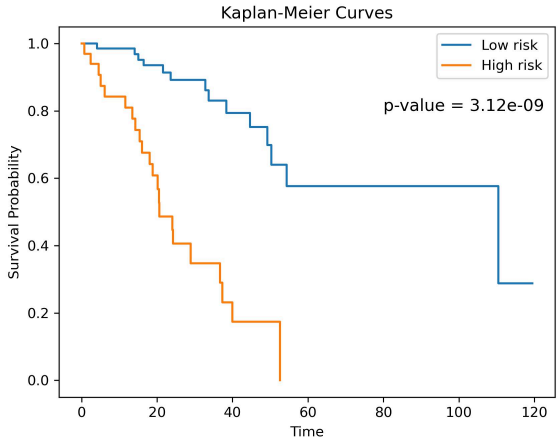

Figure 7: KM curves of the Mutation DeepSurv's risk score predictions in the testing set.

The model demonstrated high predictive accuracy for patients' risk scores when evaluated on the unseen testing set. KM survival curves revealed a clear stratification between the high-risk and

low-risk groups, with patients in the high-risk category exhibiting significantly higher mortality rates and poorer survival outcomes compared to those in the low-risk group. This distinction was further validated by a statistically significant log-rank test ($p<0.05$). These results highlight the efficacy of the DeepSurv model in using mutation data to predict patient risk scores and survival outcomes.

After training and optimization, the individual DeepSurv models achieved impressive accuracies in predicting patients' risk scores. This approach showcases the potential for integrating high-throughput molecular data with machine learning frameworks to improve accuracy and performance.

2.3 Ensemble methods

After training and optimizing, the individual DeepSurv models were integrated using different ensemble methods. Firstly, the models were combined with the average ensemble, in which the average predicted log-risk values were calculated from individual models' output. This approach was simple and efficient to implement. The average ensemble method used the predicted risk score from the different data types to increase the model's accuracy and performance. The average ensemble model had a mean c-index of 0.95 (95% CI: 0.94-0.96) and 0.78 (95% CI: 0.69-0.85) for the training and testing sets. This ensemble approach had a significantly higher predictive accuracy than any individual DeepSurv models ($p<0.05$).

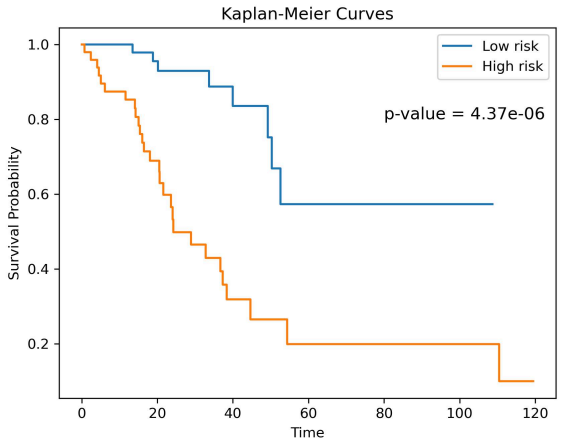

Figure 8: KM curves of the Average ensemble DeepSurv's risk score predictions in the testing set.

When applied to the unseen testing set, the average ensemble model could classify high-risk and low-risk patients efficiently. The KM curves indicated that there was a significant difference in survival of the predicted patient groups ($p<0.05$). This result suggested that by combining outputs from individual models, the average ensemble model enhanced its robustness and accuracy.

Another ensemble method employed in this project was the weighted average ensemble. While the underlying principle of the weighted average ensemble is similar to the previous ensemble method, the model combined individual models' predictions weighting each model's contribution based on its performance. Weighted average has been shown to outperform average ensemble when there is a large disparity in the individual models' performance. The mean c-index of individual DeepSurv models was used as weights to measure their performance. This approach resulted in a mean c-index of 0.95 (95% CI: 0.94-0.96) and 0.78 (95% CI: 0.71-0.84) for the training and testing sets. Weighted ensemble also increased the predictive accuracy significantly from the individual models ($p<0.05$).

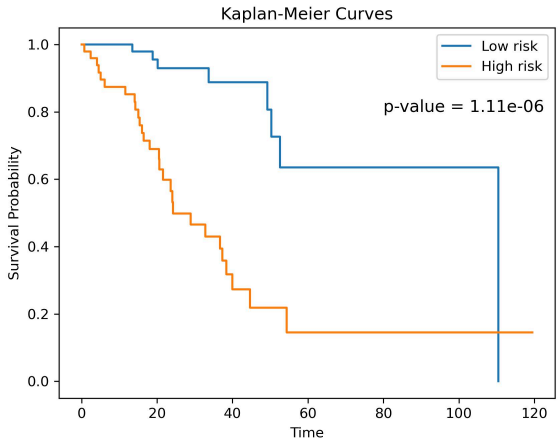

Figure 9: KM curves of the Weighted average ensemble DeepSurv's risk score predictions in the testing set.

The model had high predictive accuracy for patients' risk scores. KM survival curves showed a clear difference between the high-risk and low-risk groups, with patients in the high-risk category exhibiting higher mortality rates and poorer survival outcomes compared to those in the low-risk group. This distinction was further validated by a statistically significant log-rank test ($p<0.05$).

The stacking ensemble was also used to integrate different omics models. This approach built an artificial neural network using predictions of individual DeepSurv models as inputs. This meta-model was trained and optimized similarly to an individual DeepSurv model to estimate log-risk value with negative log partial likelihood loss function. After optimization using 5-fold cross-validation, the best-performing meta-model had 3 hidden layers and 32 filters. The stacking ensemble DeepSurv had a mean c-index of 0.95 (95% CI: 0.93-0.96) and 0.8 (95% CI: 0.73-0.87) for the training and testing sets. The meta-model also achieved a significantly higher accuracy when applied to the testing set compared to the individual models ($p<0.05$).

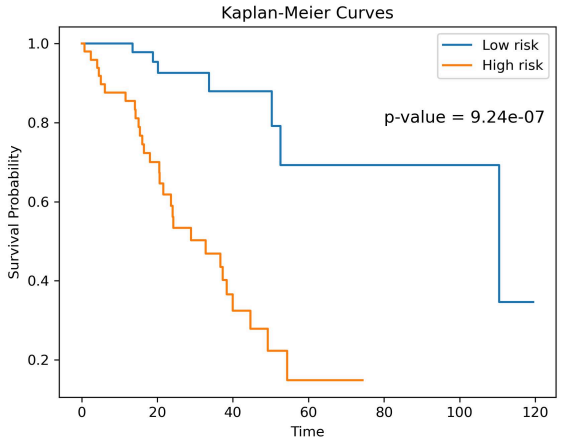

Figure 10: KM curves of the Stacking Ensemble DeepSurv's risk score predictions in the testing set.

Moreover, the stacking model could classify high-risk and low-risk patients efficiently. The KM curves indicated that there was a significant difference in survival of the predicted patient groups ($p<0.05$).

When investigating the c-index in the testing set, the ensemble methods showed a significant increase compared to the individual models ($p<0.05$). This result indicates that integrating multi-omics data could aggregate predictions, reduce error, and enhance generalization. The ensemble approaches are particularly effective in complex cancer biology where a single model might struggle to capture all underlying biological patterns.

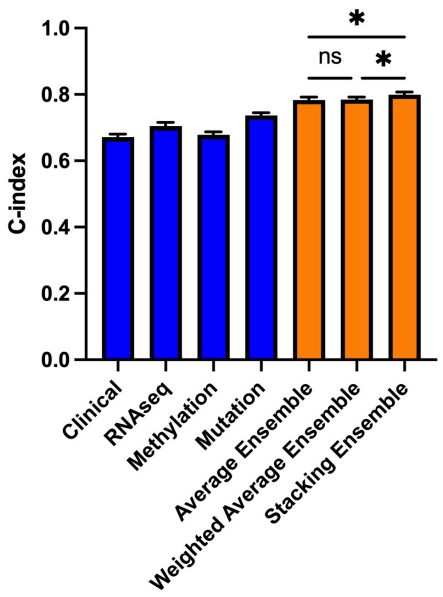

Figure 11: DeepSurv models c-index in the testing set.

Among the ensemble methods, the stacking ensemble resulted in the best-performing model ($p<0.05$), while the average and weighted average ensemble showed no significant differences. In this study, the weighted average approach did not improve from the average ensemble because the individual DeepSurv models did not exhibit great differences in their performance. However, the stacking ensemble outperformed the average ensemble because the meta-model employed a more sophisticated method to combine predictions from the base models. While averaging ensembles integrated predictions through straightforward arithmetic, stacking optimized a meta-learner to identify patterns and dependencies in the outputs of the individual DeepSurv models. This allowed the stacking ensemble to adapt to complex interactions between individual model predictions, optimizing performance for this dataset.

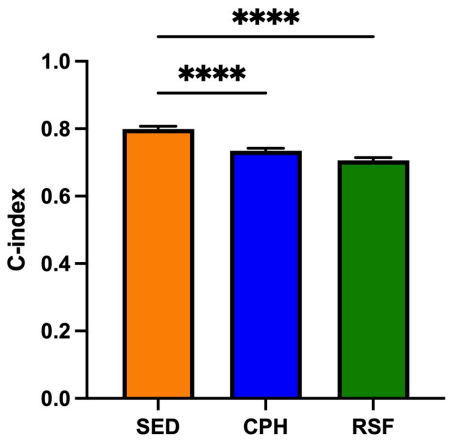

Figure 12: Performance of Stacking Ensemble Deepsurv (SED), Cox Proportional Hazard (CPH), and Random Survival Forests (RSF).

To explore whether the stacking ensemble DeepSurv (SED) model was better performing than other conventional survival analysis approaches, I compared my final SED model to Cox Proportional Hazard (CPH) and Random Survival Forests (RSF). The CPH and RSF were trained and optimized using the important features selected from different omics datasets. When applied to the testing set, the SED model had a significantly higher accuracy compared to the CPH and RSF models ($p<0.05$). This result suggested that through this project, I have developed a robust deep-learning framework that integrated multi-omics data to predict NSCLC patient survival. The integrated model outperformed all single-omics models and other traditional survival analysis methods including CPH and RSF. Among the ensemble methods, the stacking ensemble was shown to be the optimal approach for the NSCLC patient dataset.

DISCUSSION

I proposed the following timeline for this project:

1. Weeks 1-2: Data acquisition and preprocessing, including normalization, scaling, and initial exploratory analysis.

2. Week 3: Feature selection using Cox regression for DNA methylation, gene expression, mutation, and clinical data.
3. Weeks 4-5: Building and optimizing individual DeepSurv models for each dataset.
4. Weeks 6: Ensembling the individual DeepSurv models.
5. Weeks 7-8: Evaluating models' performance, report writing, and summarizing findings.

I proposed a timeline of 8 weeks to complete this project. This was a reasonable timeline and workload for the scope of this project. I finished the project according to the proposed timeline. I spent the first 2 weeks to understand and clean the data. The third week was spent on feature selection using Cox regression. I built and trained multiple DeepSurv models for each dataset. I spent week 6 performing different ensemble methods. Lastly, I evaluated all the models with the unseen testing set and wrote a project summary.

While working on this project, I have faced some challenges. The first challenge was implementing and customizing the DeepSurv model. Existing software lacked the flexibility needed for DeepSurv architecture modifications and optimization. Additionally, the default DeepSurv framework did not support training with k-fold cross-validation, which is essential for robust model evaluation. To overcome this, I dedicated significant time to developing a custom workflow that allowed me to customize DeepSurv models and incorporate hyperparameter tuning with k-fold cross-validation. Once this pipeline was in place, training and optimizing the model became efficient and straightforward. Another challenge was selecting an appropriate accuracy metric. While I initially considered using the Integrated Brier Score (IBS), which measures the mean squared differences between true outcomes and predicted probabilities, I discovered that IBS was not ideal for DeepSurv. This is because DeepSurv estimates the hazard function, which is not dependent on time, whereas IBS evaluates prediction accuracy over time. To address this, I selected the Kaplan-Meier Curve as an alternative evaluation tool due to its simplicity and effective visualization of survival probabilities.

CONCLUSION

Integrating multi-omics data for survival analysis offers a unique opportunity to capture the complex and diverse biological signals present in different molecular data types, creating accurate and comprehensive prognostic models for NSCLC. By leveraging the strengths of each omics type, including transcriptomics, DNA methylation, and somatic mutation profiles, this project aims to overcome the inherent limitations of single-omics analyses, which often fail to capture the interactions between different molecular mechanisms driving cancer progression. This project proposed a novel deep-learning framework that integrates clinical, DNA methylation, gene expression, and mutation data to predict patient survival. Univariate Cox proportional hazards regression was applied for feature selection within each omics type to identify important molecular biomarkers for NSCLC patient survival. The individual deep survival networks for each dataset were combined through multiple ensemble methods to increase the predictive accuracy and performance. The individual deep-learning models using single-omics data were able to achieve good accuracy on the testing set (clinical: 0.67, gene expression: 0.71, DNA methylation: 0.68, mutation: 0.74). However, the integration approaches significantly enhanced the performance of the deep survival models. Among the ensemble methods, the stacking ensemble DeepSurv model had the highest accuracy of 0.8, while the average and weighted average models had a significantly lower c-index. The stacking ensemble approach improved accuracy and generalizability in survival prediction by capturing non-linear, multi-dimensional interactions in multi-omics data. This integrative approach not only enhanced predictive accuracy but also increased the model's robustness, leading to more reliable and personalized survival predictions. The ensemble model's high accuracy and adaptability could have transformative implications for clinical practice. By making precise survival predictions, clinicians would be better equipped to personalize treatment plans for NSCLC patients. Furthermore, my project could set a foundation for similar applications in other types of cancer.

Future directions of this work include incorporating additional layers of omics data, such as proteomics or metabolomics, to capture even more aspects of tumor biology. These data types could reveal further biological processes and metabolic pathways contributing to cancer progression and patient prognosis. Moreover, the framework developed in this project could be adapted and applied to other cancer types or diseases where multi-omics data are available. The insights gained from such studies would enhance the understanding of cancer biology, improving survival outcomes for NSCLC patients.

REFERENCES

[1] Li C, Lei S, Ding L, Xu Y, Wu X, Wang H, Zhang Z, Gao T, Zhang Y, Li L. Global burden and trends of lung cancer incidence and mortality. Chin Med J (Engl). 2023 Jul 5;136(13):1583-1590. doi: 10.1097/CM9.0000000000002529. PMID: 37027426; PMCID: PMC10325747.

[2] American Cancer Society. Facts & Figures 2019. Accessed June 13, 2020.

[3] Yu, X., Zhao, H., Wang, R. et al. Cancer epigenetics: from laboratory studies and clinical trials to precision medicine. Cell Death Discov. 10, 28 (2024). https://doi.org/10.1038/s41420-024-01803-z

[4] Herbst, R. S., Morgensztern, D., & Boshoff, C. (2018). The biology and management of non-small cell lung cancer. Nature, 553(7689), 446–454. https://doi.org/10.1038/nature25183

[5] Katzman, J.L., Shaham, U., Cloninger, A. et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol 18, 24 (2018). https://doi.org/10.1186/s12874-018-0482-1

[6] Kim, D.W., Lee, S., Kwon, S. et al. Deep learning-based survival prediction of oral cancer patients. Sci Rep 9, 6994 (2019). https://doi.org/10.1038/s41598-019-43372-7

[7] Sasagawa Y, Inoue Y, Futagami K, Nakamura T, Maeda K, Aoki T, Fukubayashi N, Kimoto M, Mizoue T, Hoshina G. Application of deep neural survival networks to the development of risk prediction models for diabetes mellitus, hypertension, and dyslipidemia. J Hypertens. 2024 Mar 1;42(3):506-514. doi: 10.1097/HJH.0000000000003626. Epub 2023 Dec 13. PMID: 38088426; PMCID: PMC10842670.

[8] Malik, V., Kalakoti, Y. & Sundar, D. Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer. BMC Genomics 22, 214 (2021). https://doi.org/10.1186/s12864-021-07524-2