BIOS 626 Project: Machine Learning Project

Group 16: Hyeonji Ha, Seungseok, Lee, Sooyeon Oh

Submitted to: Dr. Xiang Zhou, PhD

Department of Biostatistics

University of Michigan, Ann Arbor, MI, 48104

Predicting Cancer Patient Survival Time by Using Cellular or Subcellular Resolution Spatial Proteomics Data

Abstract

Background

The aim of this project was to predict cancer patients' survival time by using cellular or subcellular resolution spatial proteomics data. Several research papers have identified the utility of spatial proteomics data in mapping call states and enhancing comprehension of tissue organisation. For example, spatial proteomics imaging techniques provide an avenue for a more intricate characterization of spatial cell patterns within tissue samples and their influence on patient outcomes.[1] In this study, various features were analyzed to enhance the prediction of cancer patients' survival duration. Different features were employed for training in the deep learning model with the aim of enhancing Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) values.

Methods

The dataset comprises features extracted from one image per patient. Each image consists of 52 channels, capturing the spatial expression pattern of 52 different proteins within the tumour tissue. The outcome variable measured is the survival time of patients, recorded in months. The training dataset includes outcome information and corresponding image data for 225 patients, while the test dataset contains image data for 56 patients.

Many features in addition to average intensity measurement of each protein on each image were extracted to improve the accuracy of the test output data. Some of the blank images were excluded as they did not contribute to further improve the accuracy. A Deep Neural Network (DNN) with 101 features was applied but no prediction improvements were observed.

After these pre-processing procedures, an ensemble model was used to improve overall performance. Instead of relying on the output of a single model, the ensemble model leverages the collective analysis of three different types of deep learning models: Convolutional Neural Network (CNN), Random Forest Regressor (RFR), and ResNet 18, and DenseNet 201.

Results

Models	***MAE***	***MSE***	***R^2***
Random Forest	37.0977	2204.3694	0.0521
CNN – RFR (denoise)	30.5553	1513.8823	0. 2229

Conclusions

The CNN-RFR (denoise) model produced the best prediction among all the models built to predict cancer patients' survival time. Because of several limitations of this study, it was difficult to predict cancer patients' survival time perfectly. There might have been confounding variables that could significantly influence the analysis of the study. Also, the scarcity of previous research on the project's topic posed challenges in selecting an appropriate model for enhanced accuracy. Despite encountering several limitations during the analysis, these findings underscore the need to optimize the model to improve predictions of cancer patients' survival time.

Introduction:

Today, despite considerable progress in early detection of a disease, cancer patients still face bleak prognoses with low survival rates. In order to solve these issues, gene expression data has been used widely since they are considered as main indicators in survival prediction of cancer.[2] However, spatial proteomics is another approach that enables the study of protein localization within cells and tissues, facilitating the understanding of underlying biological processes. This method helps to clarify the cellular heterogeneity within tissues.

Numerous researchers have spent and are currently spending extensive time and effort to find the best methodology to predict cancer patients' survival time. The development of a CNN model involves both its design and validation. Validation entails comparing the model-derived data with observations gathered from testing the predicted outcomes. Moreover, the results of this validation will indicate the accuracy of the deep learning model.

In conclusion, our project endeavors to contribute to the ongoing research on predicting cancer patients' survival time by using spatial proteomics data. By figuring out different features and utilizing them for building an appropriate CNN model, we aim to provide valuable insights that may contribute to predicting more accurate survival time of the cancer patients.

Materials and Methods:

Features

Additional features alongside the average intensity of each protein on each image were used to improve different models. The following are the 5 main feature types but not limited to:

- Statistical Features: kurtosis, entropy, standard deviation
- Morphological Features: Laplacian variance, Fourier descriptors, Wavelet features, Principal Component Analysis (PCA)
- Texture Features: contrast, correlation, Gray-level Co-occurrence Matrix (GLCM) characteristics, variations in Local Binary Pattern (LBP)
- Other Features: radial distribution features based on the center of mass, fractal dimension, Euler number, autocorrelation features, etc.

Model

Training models with extracted features

The provided training data consists of outcome and image data for 225 patients. Each image contains 52 channels, measuring the spatial expression pattern of 52 proteins on the tumor tissue. At first, 7 models were trained by extracted features of each image file: Linear Regression, Decision Tree, Random Forest, Gradient Boosting (GBM), Support Vector Machines (SVM), Multilayer Perceptron (MLP), and K-Nearest neighbors (KNN). The data for these models were split into 70% of the training and 30% of the testing partitions, without cross-validation. Lastly, K-fold cross validation was added after noticing improvements in the MAE, MSE, and R^2 values. 20% of the data were allocated for validation in each fold.

Deep Neural Network (DNN)

The DNN models were utilized to train on the extracted features, coupled with cross-validation techniques for further evaluation. Various adjustments were implemented to optimize the performance of these models. Total of 101 features were used for the final model. Initially, a

shift from MSE to MAE as the loss function was made, along with experimentation with Huber loss to handle outliers and ensure a more balanced error distribution. Additionally, hyperparameter tuning was applied, involving adjustments to lowered learning rates, increased epoch numbers, and reduced batch sizes. This was done to facilitate more refined updates and accommodate complex model structures.

Furthermore, to bolster pattern recognition capabilities, the models were deepened by adding extra layers and increasing the neuron counts in the hidden layers. Batch normalization was incorporated to stabilize learning processes. The first activation function that was used in the DNN model was ReLU. LeakyReLU and ELU activation functions were added for their ability to manage negative activations, thereby aiming to facilitate more nuanced and effective pattern recognition. Overall, these adjustments played a crucial role in refining the performance of the deep learning models for predicting cancer patients' survival time.

Image Regression CNN

The Convolutional Neural Network (CNN) modelling was structured to handle the complexities of proteomics data, particularly in predicting survival time for cancer patients. This model comprises of several key components, including convolutional layers, max pooling layers, a Squeeze-and-Excitation Block, and fully connected layers. Each component played a crucial role in extracting and processing relevant features from the input images, ultimately facilitating the prediction of continuous values such as survival time.

The initial layers of the CNN were designed to extract and downscale the raw image data while preserving important information. Following this, the Squeeze-and-Excitation block was incorporated into the network to enhance feature representation and focus the model's attention on critical features. This block adjusted the weights of each channel based on their importance, thus aiding in better feature extraction and prediction accuracy.

Furthermore, the convolutional layer delved deeper into the extracted features to uncover more complex patterns and nuances within the data. Such step allowed for a deeper understanding of the underlying characteristics of the images. Finally, the fully connected layer transformed the feature maps into a format suitable for the patients' survival time.

Random Forest Regressor (RFR)

The RFR model is an ensemble learning method designed for analyzing survival data. This is indeed one of the most important steps in this project because RFR combines multiple survival trees to perform predictions on survival time. Each tree is trained using bootstrap samples of the dataset and optimizes the risk function of survival analysis when splitting nodes.

Range of hyperparameters used were the number of trees (50, 100, 150, 200), the maximum depth of trees (none, 10, 20, 30), the minimum number of samples required to split a node (2, 5, 10), the minimum number of samples required to be at a leaf node (1, 2, 4), and the number of features to consider when splitting. Moreover, the scoring function of Mean Squared Error (MSE) is used to evaluate the performance of the model. It then moves forward to train the model on the defined hyperparameter grid using cross-validation (CV) and finds the optimal combination. In this study, 3-fold CV was used and accelerated computation using all available CPU cores.

Through the optimization process, the RFR model was able to find the optimal combination within the provided range of hyperparameters. This optimal model was then utilized

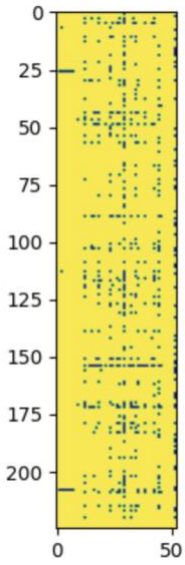
for predicting survival time of cancer patients and calculating performance metrics for MSE, MAE, and R^2 values.

Ensemble Model

Initially, all channels were combined due to a lack of understanding regarding any potential relationships between the images. However, negative R^2 values implied that each channel contained a lot of similar information. Hence, a channel wise approach was chosen to mitigate issues related to collinearity.

Three ensemble models were used: CNN – RFR, ResNet 18 – RFR, and DenseNet 201 – RFR. ResNet 18 – RFR ensemble model used vector-wise approach, whereas the CNN – RFR model used channel-wise approach. A key difference between these two ensemble models is that ResNet 18-RFR is a pretrained model specifically designed to extract features automatically from images. On the other hand, DenseNet 201 relies on dense connections between layers, allowing for maximum information flow and feature reuse. These features were then converted into vectors and processed by the RFR. On the other hand, for the CNN-RFR model, each channel's data was individually inputted into the RFR model to generate the final prediction.

Figure 1: Full Channel Lower Than Threshold Variance



Moreover, in preprocessing the data, several channels exhibited almost no meaningful information, presenting only white noise. To address this, the function for calculating channel presence was used to identify channels where the variance of pixel intensity was lower than the threshold. Then, the channels with variance lower than the threshold were identified with blue dots in the image and were imputed. This denoising technique on CNN – RFR has successfully enhanced the model prediction. While this may have caused some deletion of important features in the data, denoising ultimately improved the model prediction.

Results:

Comparing Results in 7 Different Models

Despite using different approaches to improve the DNN model, the performance did not seem to improve at all. While 101 features were used for the DNN model, only 73 features were used for the final model. This reduction has been made by removing tuple-based feature types to enhance performance. Following this adjustment, the results of the Random Forest model were considered the most effective for predicting the survival time of cancer patients. The performance of various machine learning models was assessed using three key metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and the coefficient of determination (R^2). The features were extracted without preprocessing images.

Table 1: Model Performance Using Extracted Features

Models	***MAE***	***MSE***	***R^2***
Linear Regression	379.8025	2990500.0454	-1074.1584
Decision Tree	51.36	4385.7867	-0.8715
Random Forest	37.0977	2204.3694	0.0521
GBM	37.3621	2252.5403	0.0318
SVM	37.8718	2376.3875	-0.0117
MLP	4616.5944	851274380.8464	-437461.1378
KNN	42.2542	2902.6325	-0.2576

Table 1 shows the performance of 7 different models in predicting cancer patients' survival time using extracted features. Both Linear Regression and MLP exhibit extremely poor performance with high MAE and MSE values, along with negative R^2 values. This indicates worse predictions than simply using the mean of the target variable. The MAE and MSE values for Decision Tree, SVM, and KNN do not differ significantly from other models but were considered inefficient due to the negative R^2 values.

Only the two models, Random Forest and GBM, seem to show positive R^2 values for predicting cancer patients' survival time. Among the two, Random Forest demonstrates better performance with an R^2 value of 0.0521, indicating that 5.21% explains the model. Overall, while some models maintain consistent performance across evaluation methods, others exhibit notable discrepancies, suggesting potential variability in their generalization capabilities.

Table 2: Ensemble Models Performance

Models	***MAE***	***MSE***	***R^2***
DenseNet 201 – RFR	36.9255	2314.1462	0.0274
ResNet 18 – RFR	35.0185	2086.2312	0.1232
CNN – RFR	30.8689	1633.0499	0.1617
CNN – RFR (denoise)	30.5553	1513.8823	0.2229

The following results indicate the performance of three ensemble models in predicting cancer patients' survival time. The values of MAE, MSE, and R^2 are all better in the CNN-RFR model among the three models. At first, R^2 value for ResNet 18 – RFR was higher than that for

DenseNet 201 – RFR because of its complexity. This result implies a potential issue with vanishing gradients that can hinder training effectiveness. The R^2 value of the CNN-RFR model is the highest at 0.1617, suggesting that it explains around 16.17% of the variance in the data. In order to improve the model further, a denoising technique has been applied. This then significantly increased an R^2 value. Overall, while these three models show some ability to predict cancer patients' survival time, the CNN-RFR (denoise) model performs the best among them in terms of lower error metrics and higher explained variance.

Discussions:

The goal of our study was to explore various methods to accurately predict the survival time of cancer patients. To achieve the best prediction outcomes, future research may need to explore additional biological factors or protein types that could yield meaningful features. For example, incorporating features such as cell shapes or sizes could prove beneficial, especially when leveraging cellular or subcellular resolution spatial proteomics data. However, it is essential to acknowledge the potential presence of confounding variables. Failure to account for such can lead to biased predictions and inaccurate assessments of the performance of the deep learning models.

Moreover, the limited amount of previous research on using spatial proteomics poses difficulties in choosing the most suitable model to improve accuracy. The lack of existing literature reviews creates a gap in established methodologies, which made it difficult to find the most effective approach for our research question. The scarcity not only complicates the initial selection process but also hampers the validation and refinement stages, as there are limited precedents to draw upon for calibration.

The presence of negative R^2 values of the models raises concerns about the applicability of the results. While negative R^2 value may seem abnormal, they simply indicate the model's poor performance on predicting cancer patients' survival time. As such, this can be interpreted that the model's performance is even lower than a simplest baseline model. This problem is not necessarily a problem with a code itself. Instead, negative R^2 values strongly signal that the model poorly explains the data, leading to the exclusion of these models from final consideration.

Lastly, the limitation of insufficient sample size warrants consideration in our study. The relatively small size of the sample population may have restricted the robustness of our findings and the generalizability of our conclusions. With a larger and more diverse sample, it would have been possible to capture a broader range of variability in the data, potentially leading to more reliable and representative results. A larger sample size would have allowed for more detailed feature extraction and accurate results. Additionally, the lack of computational resources posed challenges in processing and analyzing the data efficiently. Nevertheless, our findings emphasize the importance of optimizing our model to enhance predictions of cancer patients' survival time.

References

[1] Dayao, M. T., Trevino, A., Kim, H., Ruffalo, M., D'Angio, H. B., Preska, R., Duvvuri, U., Mayer, A. T., & Bar-Joseph, Z. (2023). Deriving spatial features from in situ proteomics imaging to enhance cancer survival analysis. *Bioinformatics (Oxford, England)*, *39*(39 Suppl 1), i140–i148. https://doi.org/10.1093/bioinformatics/btad245

[2] Bashiri, A., Ghazisaeedi, M., Safdari, R., Shahmoradi, L., & Ehtesham, H. (2017). Improving the Prediction of Survival in Cancer Patients by Using Machine Learning Techniques: Experience of Gene Expression Data: A Narrative Review. *Iranian journal of public health*, *46*(2), 165–172.

[3] Lualdi, M., & Fasano, M. (2021). Features Selection and Extraction in Statistical Analysis of Proteomics Datasets. *Methods in molecular biology (Clifton, N.J.)*, *2361*, 143–159. https://doi.org/10.1007/978-1-0716-1641-3_9