Title: "A Machine Learning Approach to Proteomic Based Breast Cancer Classification"

Author: Kelsey Nicholson

Abstract:

Breast cancer is one of the most prevalent types of cancer among women worldwide. It is a complex disease that can be categorized into different molecular subtypes, HER2, Luminal A, Luminal B, and Basal-like. Understanding the unique cellular profiles of each subtype can lead to the development of more effective and specialized treatments. Currently, molecular subtype classification is based on genomic sequencing technologies. However, proteomic data has been shown to have low correlation with genomic data, indicating that genomic variations are not fully translated at the protein level. This project sought to use proteomic data obtained from the Proteomic Data Commons to classify data into one of the four molecular subtypes. After preprocessing the data and detecting outliers, the dataset was split into training, validation, and testing sets. Four classification models, Support Vector Machine (SVM), logistic regression, Feedforward Neural Network (FNN), and 1D Convolutional Neural Network (CNN) were trained and tuned using Randomized Search CV or keras tuner RandomSearch. The models were evaluated based on their final macro f1 scores on the training and validation scores, the best models were selected to be run on testing data. The highest score achieved by SVM and Logistic Regression models during training were .660 and .664 The FNN and CNN models performed better in training with f1 scores of .75 and .82. These models were selected for final testing and achieved scores of .699 and .880 on unseen testing data. These results demonstrate the significant performance of the 1D CNN model on the dataset. This project shows the importance of analyzing differences of cellular function at the protein level to create more targeted treatment options for subtype-specific breast cancer classifications.

Introduction:

In recent years, precision medicine and gene therapy have played a significant role in treating cancer, emphasizing personalized treatments based on things like cellular profile of a disease and the patient's genome. Breast cancer is typically classified into four molecular subtypes based on gene expression: HER2, Luminal A, Luminal B, and Basal-like. Molecular subtype classification leads to different standards of care and can improve treatment for patients. For example, Luminal A cancers tend to grow more slowly than other cancers and have a good prognosis, while HER2-enriched cancers tend to grow faster than luminal cancers and can have a worse prognosis (Sheng 2022). However, HER2-enriched is usually successfully treated with targeted therapy medicines aimed at the HER2 protein.

Proteomic data has been shown to have low correlation with genomic data, indicating that genomic variations are not fully translated to the protein level. This makes it essential to analyze differences of cellular function at the protein level, in addition to genomic data, to create more targeted treatment options for subtype-specific breast cancer classifications.

Understanding the unique cellular profiles of breast cancer subtypes can lead to the development of more effective treatments that cater to the individual needs of patients.

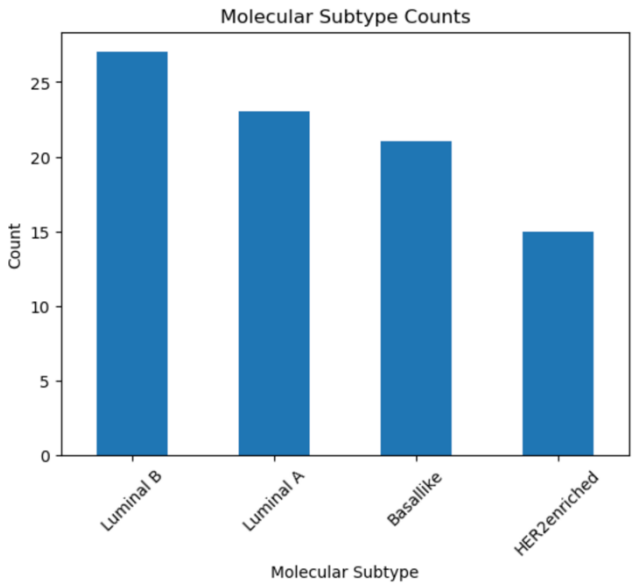
Background:

With the prevalence of precision medicine and gene therapy, breast cancer has been studied extensively at the genomic level. Currently, the subtypes have been classified with genomic sequencing technologies, but in recent years advances in mass spec technologies have provided deep proteome coverage and data. For example, the study in 2017, "Proteomic maps of breast cancer subtypes," found that in a comparison between proteomic data and genomic data, there is low correlation between the copy number variants in the genome and the relative change at the protein levels. This indicates that genomic variations are not translated or only partially translated to the protein level. This is important because germline copy number variations at the genomic level are associated with breast cancer risk and prognosis. Thus, it could be very useful to analyze differences of cellular function at the protein level.

Methods and Results:

The data for this project was obtained from the Proteomic Data Commons, which contains proteomic and clinical datasets for the four molecular subtypes. The proteomic dataset contains 12,000 genes and corresponding protein ratios for each patient, while the clinical dataset contains the molecular subtypes for each patient. These datasets were combined on the patients' unique IDs from the proteome and clinical datasets. After obtaining the data, the first step was to merge the two datasets and start preprocessing the data. This involved dropping unnecessary columns and renaming the remaining columns to match the clinical dataset. To standardize the data and streamline the analysis process, all data values were converted to floats. Once the data was cleaned and processed a histogram was used to show the class breakdowns to gain insight into the data before splitting. The molecular subtype counts were plotted to visualize the breakdown of each class (Figure 1). This plot indicated that the dataset has an imbalanced class breakdown and would need to be considered for downstream model optimization and evaluation. Outliers were detected using quartile limits and columns were dropped with significant outliers, this brough the data from 12,551 features to 3,786. The remaining data was shuffled and split into training, validation, and test sets using the train_test_split function from the scikit-learn. This function used an 80/20 split

Figure 1:



with the remaining .80 of the training data further split into specific training and validation sets. To decide how to handle NA values, a distribution plot of the training data was created (Figure 2). After analyzing the data's distribution and accounting for missing values, it was decided to use data imputation using the median value. This decision was made because of the data's decent range of values, even after removing outliers. The training data was scaled, and the subtype column classes were encoded to integers using the label encoder from scikit-learn. After data preprocessing two models were selected for classification, Support Vector Machine (SVM) and logistic regression.

The hyperparameters for the SVM and Logistic Regression models were tuned using Randomized Search CV and 5-fold cross-validation. The best hyperparameters were selected based on macro f1 score. The final SVM and Logistic Regression models were trained on the training data using the best hyperparameters obtained from hyperparameter tuning. The model was then evaluated using validation data. After preprocessing, the SVM model achieved a final f1 score of 0.660, while the Logistic Regression model scored 0.638. These results highlight that SVM model may perform slightly better on the data (Table 1).

Figure 2:

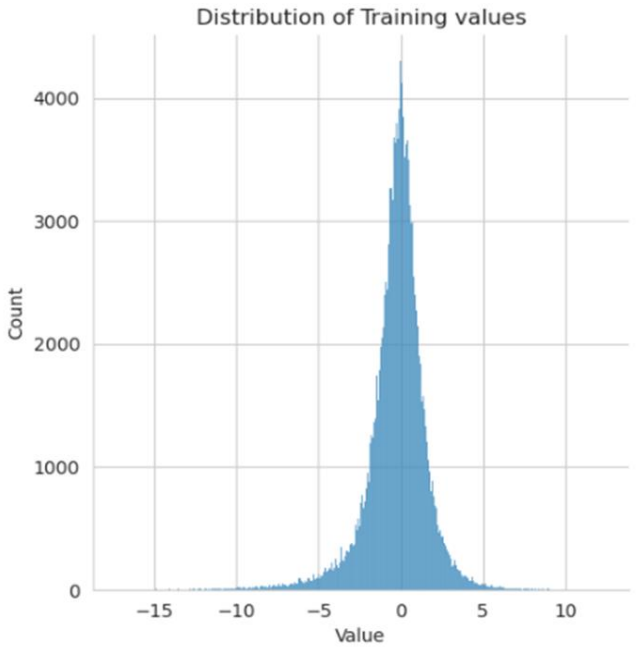


Table 1:

Model Type	Macro F1 Score on Training Data
SVM all Features	0.660
Logistic Regression all Features	0.638
SVM with Lasso Feature selection	0.606
Logistic Regression with Lasso Feature selection	0.664
SVM with PCA for feature selection (10 components)	0.679
Logistic Regression with PCA (10 components) for feature selection	0.545
FNN all Features	**0.750**
CNN all Features	**0.821**
FNN with Lasso Selected Features	**0.714**
CNN with Lasso Selected Features	**0.720**

To improve the models' performance, the first feature selection method used was the Lasso regularization technique. After applying Lasso regularization techniques, the Logistic Regression model showed a slight improvement in performance with score of 0.606, however the SVM model did not perform as well after feature selection, with a decrease in performance and score of .606 (Table 1). The next approach for feature selection was to apply principal component analysis (PCA), this technique identifies the most informative components in a dataset. The explained variance ratio of each principal component was plotted to determine the appropriate number of components (Figure 3). Based on the variance plot, 2, 5, and 10 components were used. Different data reduction techniques were employed and an unsupervised K-means clustering approach was used as a component of exploratory analysis. The project iterated over 2, 5, and 10 components to determine the optimal number of components for dimensionality reduction. The SVM and Logistic Regression models with the highest f1 score used 10 components (Table 2). Additionally, the purity and silhouette scores were calculated to evaluate the effectiveness of the K-means clustering algorithm on unlabeled data. The purity score measures the proportion of correctly classified data points in the same cluster, while the silhouette score measures the quality of the clusters.

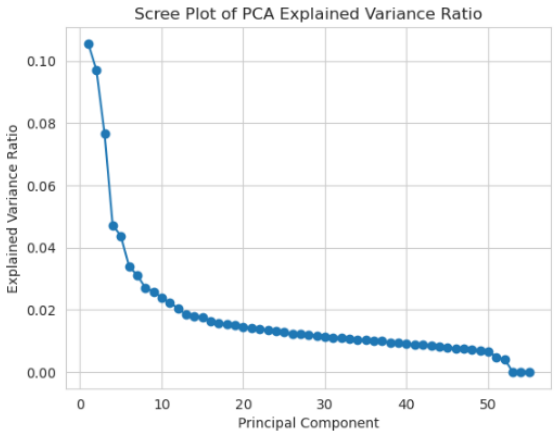


Table 2:

	2 components	*5 components*	*10 components*
PCA			
F1-score (logistic regression)	0.425	0.517	0.679
F1-score (SVM)	0.250	0.587	0.545
Purity score	0.436	0.491	0.527
Silhouette score	0.364	0.209	0.144
UMAP			
F1-score (logistic regression)	0.396	0.296	0.562
F1-score (SVM)	0.292	0.483	0.625
Purity score	0.527	0.527	0.509
Silhouette score	0.411	0.354	0.346
Kernel PCA			
F1-score (logistic regression)	0.532	0.440	0.549
F1-score (SVM)	0.250	0.440	0.375
Purity score	0.436	0.418	0.436
Silhouette score	0.377	0.311	0.148

The K-means clustering algorithm produced a purity score of approximately .5 and a silhouette score of approximately .4. These scores were compared to the actual classes to evaluate the effectiveness of the unsupervised method in clustering the data

into similar classes. The results showed decent clustering, indicating that the K-means algorithm was able to group similar data points together, but could still use improvements, indicating this may not be the best method. To further test the classification f1 score, dimensionality reduction techniques such as UMAP and Kernel PCA were also utilized. The UMAP and Kernel PCA transformed data was then evaluated using f1 score, with the results shown in Table 2. The Logistic Regression model using PCA for data reduction achieved the highest f1 score, indicating this may be a good model for classification. Overall, the combination of supervised and unsupervised methods, as well as dimensionality reduction techniques, allowed for a thorough evaluation of the data classification.

After identifying several potential candidates for final model evaluation, the approach was expanded to include the creation of two deep learning models: a Feedforward Neural Network (FNN) model and a 1D Convolutional Neural Network (CNN) model. To construct the FNN model, a function was created to model the FNN architecture and optimize the number of hidden layers, the number of units, optimizer learning rate, and the dropout rate. Also, a custom function was created to calculate the macro f1 score. The keras tuner with RandomSearch method was utilized to search for the optimal hyperparameters, and the best model was trained. Once the optimal hyperparameters were found, the best model was trained on the training data and evaluated on the validation data. The final score on the training data was .75. To construct the CNN model, a function was created to model the architecture of a CNN and optimize the number of filters, units in dense layer, optimizer learning rate, and kernel size for the first convolutional layer. The data was also reshaped to fit the Conv1D layer. The keras tuner with RandomSearch method was utilized to search for the optimal hyperparameters for the CNN model. Once the optimal hyperparameters were found, the best model was trained on the training data and evaluated on the validation data with an f1 score of .82. Since these models produced the highest scores during training, lasso feature selection was also executed to see if model scores would improve further. However, both scores decreased, but still outperformed the SVM and Logistic Regression models.

In the final stage of the project, the performance of both the FNN and CNN models were selected as the highest performing models and evaluated on the testing set. It was observed that the CNN model outperformed the FNN model, achieving a higher macro f1 score on all features with a score of 0.88 as shown in Table 3. Additionally, when Lasso feature selection was employed prior to running the models, the FNN model returned an improved f1 score of 0.81, while the CNN model's f1 score decreased to 0.68. These results indicate that the CNN model is more effective at utilizing the available features and extracting useful information. Overall, this project demonstrated that deep learning models, such as the FNN and CNN models, are promising candidates for utilizing proteomic data for molecular subtype classification.

Table 3:

Model Type	Macro F1 Score on Testing Data
FNN all Features	0.699

CNN all Features	0.880
FNN with Lasso Selected Features	0.810
CNN with Lasso Selected Features	0.681

Discussion:

The results presented in this project show that a combination of machine learning models, feature selection methods, unsupervised clustering techniques, and deep learning models can be used to effectively classify molecular subtypes using proteomic data. The first part of the analysis involved preprocessing the data and applying two classification models, Support Vector Machine (SVM) and Logistic Regression, to predict molecular subtype. Outliers were detected using quartile limits, and columns with significant outliers were dropped. This reduced the number of features from 12,551 to 3,786. The final SVM and Logistic Regression models were trained on the training data using the best hyperparameters obtained from hyperparameter tuning. The models were evaluated using validation data, and the SVM model achieved a final f1 score of 0.60, while the Logistic Regression model scored 0.638.

The Lasso regularization method brough the features from 3,786 to 124 and was effective in improving the performance of the Logistic Regression model, although it had a negative impact on the SVM model. This suggests that the relevance of the selected features may differ between classification algorithms. PCA was employed to identify the most informative components in the dataset, and K-means clustering was used to explore the data and evaluate the effectiveness of unsupervised clustering as an option for classification. The results showed that the K-means algorithm was able to group similar data points effectively, although there was still room for improvement in terms of clustering quality. Thus, this method was not used for testing.

The application of deep learning models, specifically the FNN and CNN models, showed the most promising results. The use of keras tuner allowed for efficient hyperparameter tuning and helped to avoid overfitting of the models. The CNN model outperformed the FNN model in all features, achieving a higher macro f1 score on the test set. Additionally, the results showed that Lasso feature selection improved the performance of the FNN model, while decreasing the performance of the CNN model. These findings suggest that deep learning models are promising candidates for the classification of molecular subtypes using proteomic data. However, there is still room for improvement in terms of feature selection and model type. Future studies could explore alternative feature selection methods. Overall, the approach presented in this project provides a framework for the classification of molecular subtypes using proteomic data, which could have important implications for personalized medicine and cancer diagnosis.

Conclusion:

In this project, the performance of SVM, logistic regression, and deep learning models were evaluated for classification of breast cancer molecular subtype. While the results indicated promising potential for the use of proteomic data and deep learning in the classification of molecular subtypes, the small dataset, high dimensionality of features, and imbalanced class distribution remained significant limitations. Nonetheless, the SVM and logistic regression models, along with the FNN and CNN models, demonstrated moderate to high f1 scores on testing data. Overall, this project provides useful insights into the application of deep learning models in classification tasks and could potentially inform the development of personalized medicine and cancer diagnosis.

References:

Sheng, J., MD (2022, November 8). *Molecular Subtypes of Breast Cancer*. BreastCancer. https://www.breastcancer.org/types/molecular-subtypes

Tyanova, S. (2016, January 4). *Proteomic maps of breast cancer subtypes*. Nature Communications. https://www.nature.com/articles/ncomms10259

(2023, January 12). *Key Statistics for Breast Cancer*. American Cancer Society. cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html