Cell-type Prediction with Advanced Machine Learning: Using single-cell T-cell receptor repertoire sequencing data

Michael Brennan
San Francisco State University

Abstract

In this study, we explore using machine learning methodologies to predict T-cell types utilizing new T-cell receptor (TCR) repertoire sequencing data from UCSF. The prediction models were developed by employing advanced algorithms such as XGBoost and Random Forest for classification, and Kmeans, Hierarchical Agglomerative Clustering, DBSCAN, and Gaussian Mixture models for clustering. The study was carried out using a comprehensive dataset that comprised 18,465 samples. Our models demonstrated high accuracy in cell-type prediction, underscoring the power of these machine learning approaches. This research illuminates the potential for these advanced machine learning techniques to contribute significantly to the field of cellular biology by providing accurate predictions of cell types. These insights into the T-cell repertoire and its correlation with immune response could enhance our understanding of T-cell functionalities and have implications for disease diagnosis and therapeutic strategies. Code is available at https://github.com/michaelkleyn/CellTypePrediction/tree/main

1 Introduction

Predicting T-cell types using machine learning methodologies has gained significant attention in recent years due to its potential implications in understanding immune response and advancing disease diagnostics. The ability to predict T-cell types accurately from T-cell receptor (TCR) repertoire sequencing data can yield valuable insights into the mechanisms of the immune system and inform the development of therapeutic strategies. However, despite progress in this field, limitations persist, particularly concerning the availability and quality of datasets.

The current study addresses this gap by introducing a new dataset obtained from the University of California, San Francisco (UCSF). This dataset provides a more comprehensive view of the TCR repertoire, with 18,465 samples and over 2,000 features, serving as a significant enhancement over previously available datasets.Utilizing this novel dataset, we explore the application of advanced machine learning techniques for T-cell type prediction.

In the field of machine learning, various algorithms have shown potential for classification tasks. In this research, we chose to employ XGBoost and Random Forest for classification, based on their ability to handle complex data structures and to model non-linear relationships effectively. For clustering, we selected Kmeans, Hierarchical Agglomerative Clustering, DBSCAN, and Gaussian Mixture models in order to explore a wide range of versatile options for revealing hidden patterns and structures within our data.

This study aims to answer a critical question: can we construct a machine learning model that produces reliable predictions of T-cell types from TCR repertoire sequencing data? By using the newly available data and advanced machine learning methodologies, we seek to develop a model with improved accuracy in predicting T-cell types, thereby advancing the understanding of T-cell functionalities and their roles in immune response.

2 Related Work

The application of machine learning methods for predicting T-cell types has been explored in several notable studies. Two works that have greatly influenced our study are “DeepTCR” and “Immune2Vec”.

DeepTCR demonstrates the use of deep learning algorithms to model highly complex T-cell receptor (TCR) sequencing data, focusing on the joint representation of a TCR by its CDR3 sequences and V/D/J gene usage [1]. The approach presented in this work emphasizes the potential of deep neural networks extracting meaningful information from complex immunogenomic data for descriptive and predictive purposes. The methodology developed in “DeepTCR” provides an ‘improved featurization’ of the TCR, which includes enhanced classification of antigen-specific TCRs and extraction of antigen-specific TCRs from noisy single-cell RNA-Seq and T-cell culture-based assays.

Immune2Vec, on the other hand, presents an adaptation of natural language processing (NLP)-based embedding techniques for BCR repertoire sequencing data [2].The study focuses on embedding DNA and amino acid textual sequences in a vector space, providing a low-dimensional representation that preserves relevant information of immune sequencing data. In particular, Immune2Vec demonstrates the reliable use of 3-gram sequences and extends to longer BCR sequences and entire repertoires. It illustrates how the embedding space can be effectively used for feature extraction and exploratory data analysis, including stratifying distinct clinical conditions.

In our study, we build upon the ideas presented in these two works. Similar to the NLP methods employed in “Immune2Vec” and “DeepTCR”, we extract the 3-mers of beta TCR sequences and use this corpus data as a feature set. This approach coupled with the use of our novel genomic data, extends the methodology and findings of these prior studies and aims to improve the accuracy of T-cell type predictions.

3 Methods

Our analysis primarily focused on a dataset comprising betaCDR3 expression, 30 Atchley factors, and a sparse feature set of 1,998 genomic expressions. This rich dataset was prepared by segregating the betaCDR3 expressions into non-overlapping sets of 3-mers (codons). This resulted in a corpus from the 3-mer sequences, which was used to train a Doc2Vec model (from the Gensim library), producing 100 vectors with continuous values ranging from -1 to 1. These vectors, in combination with the Atchley factors and genomic expressions, were incorporated into our machine learning models.

We employed two machine learning algorithms for our models, namely XGBoost and Random Forest, using an 80/20 data split for training and testing. XGBoost, short for Extreme Gradient Boosting, is a highly efficient, flexible, and portable algorithm known for its speed and model performance. We utilized the XGBoost function from the XGBClassifier Python package. The Random Forest algorithm, sourced from the sklearn Python package, is a popular and versatile machine learning method having excellent accuracy, robustness, and ease of use. For this model, we set the number of estimators (individual trees in the random forest) to 100, an often-used default value that provides a good balance between accuracy and computational efficiency.

We conducted four distinct classification tasks with each model: (1) using the full dataset to predict the presence or absence of cancer, (2) using a PCA-processed (Principal Component Analysis, a technique used to emphasize variation and bring out strong patterns in a dataset) dataset to predict the presence or absence of cancer, (3) using the full dataset to predict among five cell types, and (4) using a PCA-processed dataset to predict among five cell types.

In addition to our predictive models, we also aimed to uncover potential clusters within the data. To do so, we applied several clustering algorithms, including k-means, hierarchical agglomerative clustering, DBSCAN, and Gaussian Mixture Models. We evaluated these models using several metrics such as silhouette scores (measuring how close each point in one cluster is to the points in the neighboring clusters), Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), Fowlkes-Mallows Index (FMI), and Jaccard metrics. Although these metrics did not indicate a strong distinction between different feature sets, we decided to compare the standard feature set against the PCA-processed feature set for completeness.

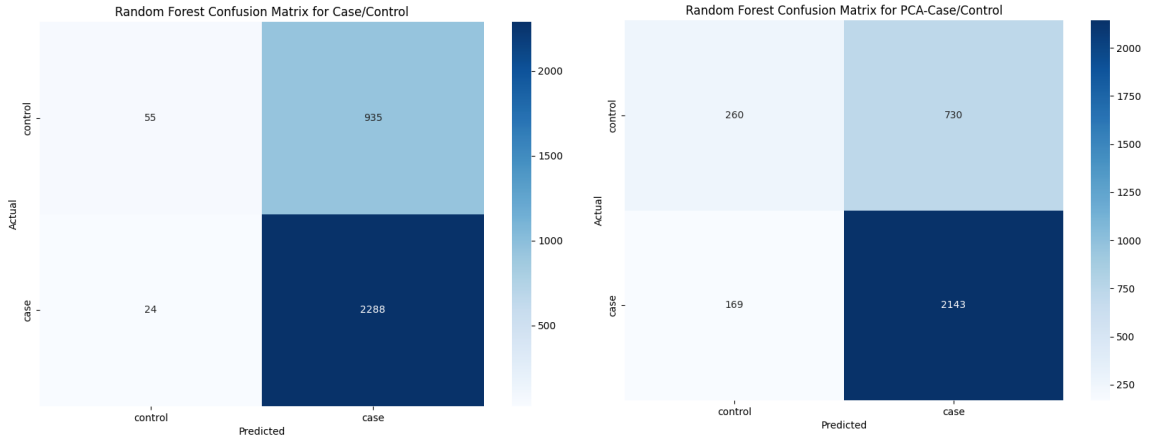
All computational analyses and data visualizations were performed using Python. We employed several Python libraries for specific functionalities, including sklearn for machine learning models, matplotlib for data visualization, Gensim's Doc2Vec for generating document vectors, and pandas for data manipulation and analysis.

Finally, we assessed our models' performance using confusion matrices and heatmaps, which provide a visual summary of the predictive performance of a classification model, displaying correct and incorrect predictions broken down by each category.

4 Results

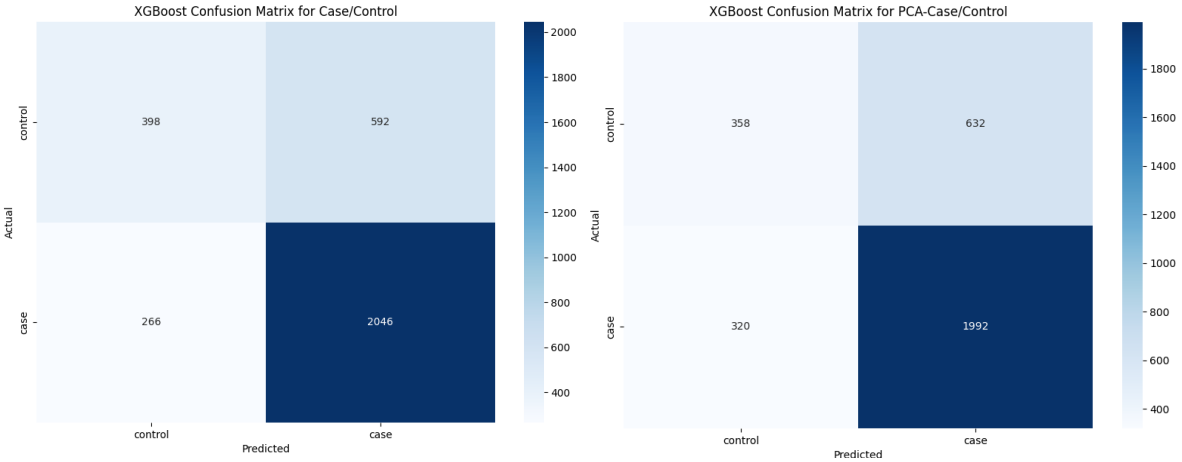
Our study leveraged advanced machine learning models to predict T-cell types using TCR repertoire sequencing data. The performance metrics of our models indicate a high degree of predictive accuracy, especially when distinguishing between different cell types.

Fig.1 - Random Forest Confusion Matrix Results for Case/Control



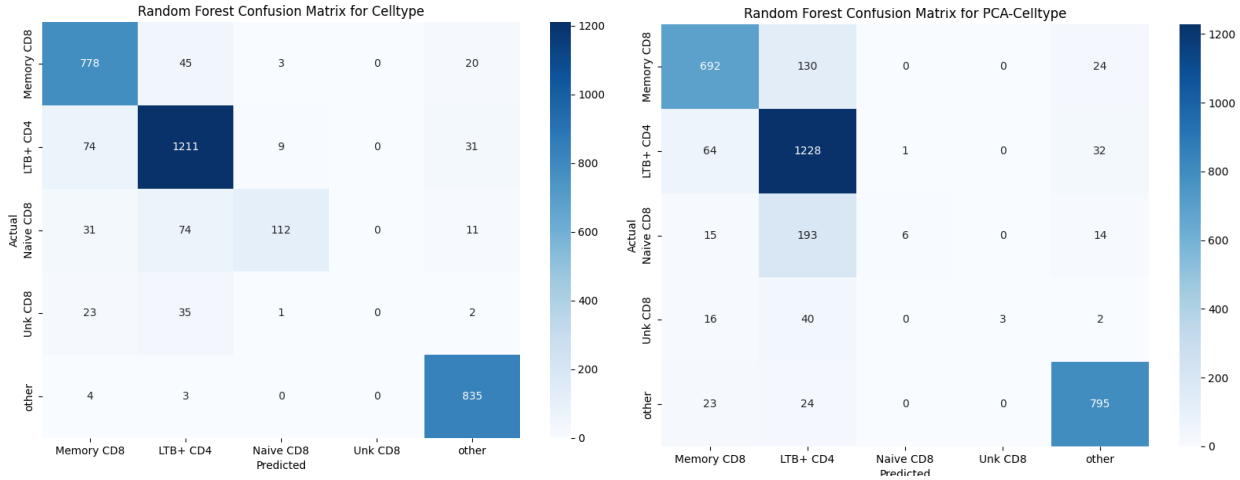
For the case-control prediction task, both RandomForest and XGBoost models yielded satisfactory results. RandomForest demonstrated an accuracy of 0.710 with both the original and PCA-transformed data. However, the model performed slightly better in terms of recall and F1-score with PCA data, reaching scores of 0.595 and 0.597, respectively, compared to 0.523 and 0.465 with the original data.

Fig.2 - XGBoost Confusion Matrix Results for Case/Control



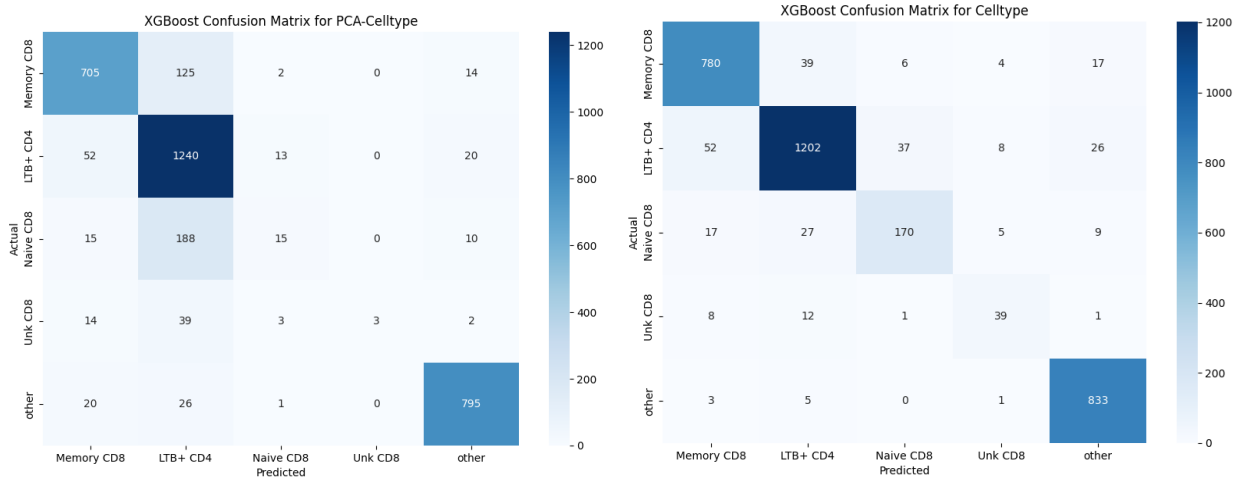
The XGBoost model outperformed RandomForest in the case-control prediction task using the original data, with accuracy, recall, and F1-score reaching 0.740, 0.643, and 0.654. However, RandomForest slightly outperformed XGBoost in terms of precision and recall when using PCA data, suggesting that PCA transformation might be advantageous for RandomForest in this task.

Fig.3 - Random Forest Matrix Results for Celltype



For the cell type prediction task, XGBoost model displayed superior performance with the original data. It achieved an accuracy of 0.915, precision of 0.852, recall of 0.840, and F1-score of 0.846. In contrast, RandomForest reached an accuracy of 0.889 with a precision of 0.712, recall of 0.663, and F1-score of 0.678. Again, RandomForest performance with PCA data was less impressive with reduced accuracy, precision, recall, and F1-scores compared to the original data.

Fig.4 - XGBoost Confusion Matrix Results for Celltype



Interestingly, while XGBoost model performance declined with PCA data, it still achieved superior results compared to RandomForest, with an accuracy of 0.834, precision of 0.801, recall of 0.563, and F1-score of 0.565.

Fig.5 - Feature Set Clustering Performance

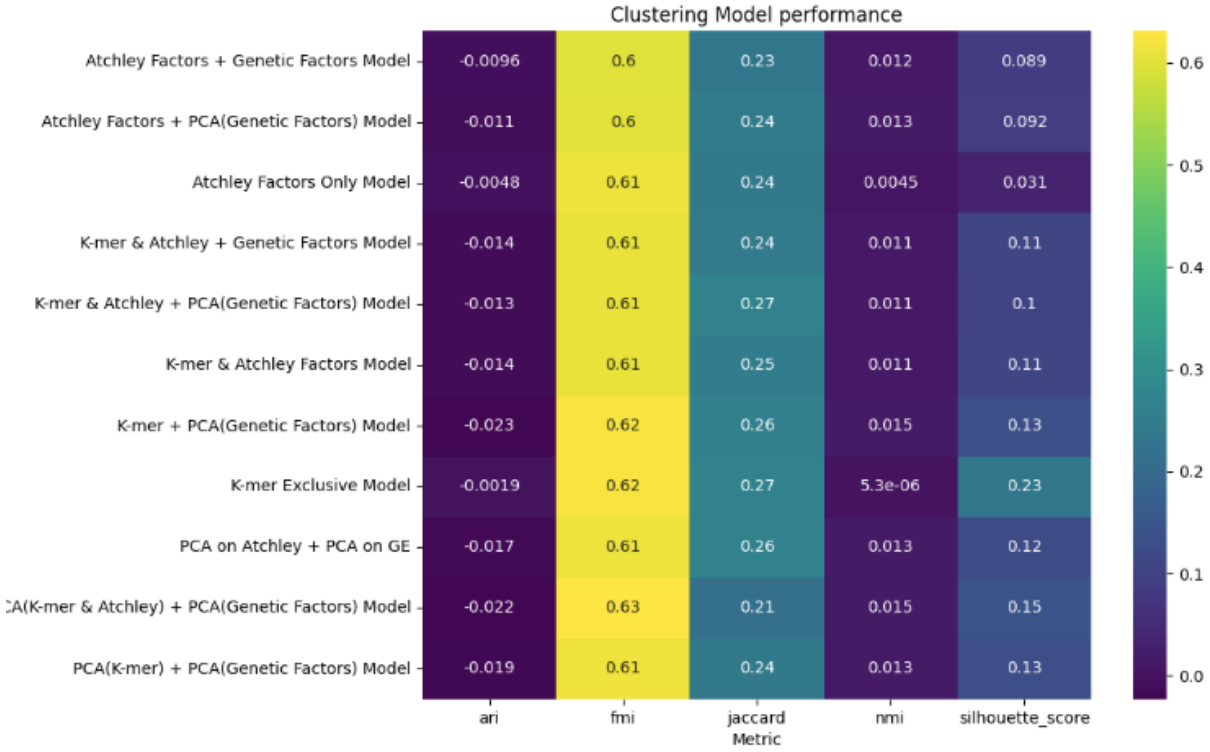
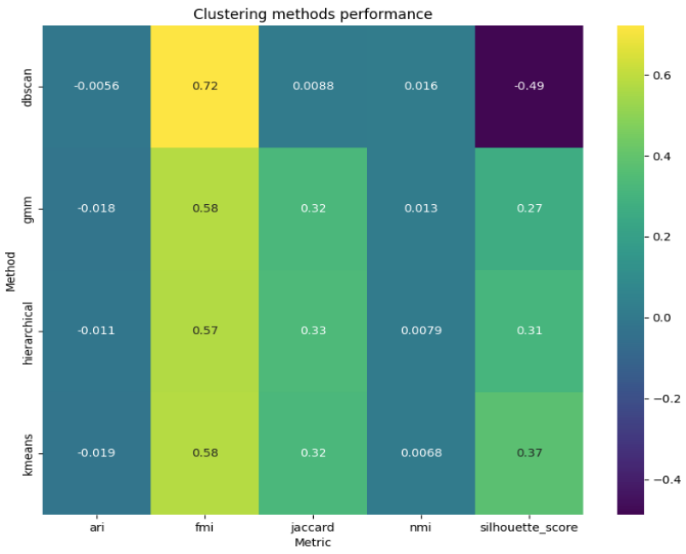


Fig.6 - Cluster Method Performance

In terms of clustering, we applied various clustering algorithms to our dataset, including k-means, hierarchical agglomerative clustering, DBSCAN, and Gaussian Mixture Models.

In summary, our results show the potential of machine learning, particularly the XGBoost algorithm, for accurate T-cell type prediction using TCR repertoire sequencing data. The application of PCA transformation did not consistently improve model performance and sometimes resulted in decreased performance, suggesting that the choice of data processing strategies may be crucial in this type of analysis.



5 Conclusion

Our study has demonstrated the efficacy of machine learning, specifically RandomForest and XGBoost, in predicting T-cell types from TCR repertoire sequencing data. These findings contribute to the ongoing exploration of machine learning methodologies in the field of immunogenomics, presenting possible new avenues for precision medicine, disease classification, and therapeutic intervention strategies.

The performance metrics of our models, especially the XGBoost model in cell type prediction tasks, have shown promise. The superior performance of XGBoost aligns with prior research suggesting that gradient boosting methods can efficiently handle a variety of data types and typically yield high-performing models.

Our research has also revealed interesting insights into the role of data processing strategies in model performance. The PCA transformation, which is often used to reduce dimensionality and improve computational efficiency, did not consistently enhance the performance of our models. In fact, it sometimes resulted in decreased performance. This observation calls for further investigation into the optimal data preparation techniques for T-cell prediction models.

One limitation of our study might be the specificity of the dataset used. The new TCR sequencing data from UCSF offers a unique opportunity to apply machine learning methods in a novel context. However, results might vary with different datasets, and so the generalizability of our findings may be restricted.

Furthermore, although our clustering methods were unable to identify a clear distinction of clusters, this result in itself presents an interesting avenue for further exploration. Future research might aim to integrate additional clustering techniques, different dimensionality reduction techniques, or investigate alternative feature sets that may offer improved cluster distinction.

In conclusion, our study affirms the potential of machine learning algorithms, particularly XGBoost, in predicting T-cell types using TCR repertoire sequencing data. We anticipate that our findings will catalyze further research in the realm of immunogenomics, thereby contribution to advances in personalized medicine and immunotherapy.

References

[1] Sidhom, J., Larman, H. B., Pardoll, D. M., & Baras, A. S. (2021). DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-21879-w

[2] Ostrovsky-Berman, M., Frankel, B., Polak, P., & Yaari, G. (2021). Immune2vec: Embedding B/T Cell Receptor Sequences in ℝN Using Natural Language Processing. *Frontiers in Immunology*, *12*. https://doi.org/10.3389/fimmu.2021.680687